TRANSPORTATION RESEARCH

RECORD

No. 1408

Highway Operations, Capacity, and Traffic Control

Intelligent Vehicle Highway Systems

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1993

Transportation Research Record 1408
ISSN 0361-1981
ISBN 0-309-05555-5
Price $28.00

Subscriber Category
IVA highway operations, capacity, and traffic control

TRB Publications Staff
Director of Reports and Editorial Services: Nancy A. Ackerman
Associate Editor/Supervisor: Luanne Crayton
Associate Editors: Naomi Kassabian, Alison G. Tobias
Assistant Editors: Susan E. G. Brown, Norman Solomon
Production Coordinator: Sharada Gilkey
Graphics Coordinator: Terri Wayne
Office Manager: Phyllis D. Barber
Senior Production Assistant: Betty L. Hawkins

Printed in the United States of America

Transportation Research Record 1408

Contents

Foreword

The papers in this Record represent results of research studies focusing on intelligent vehicle highway systems (IVHS).

Brand discusses predictive models to evaluate IVHS improvements, including the formulation of inputs that make it possible to anticipate the important consequences of IVHS and therefore carry out benefit-cost analysis of new investments as well as collect the appropriate data for planning and evaluating operational field tests.

Cheslow and Hatcher identify and evaluate five alternative architectures for advanced traffic management (ATM) and advance traveler information systems (ATIS). These activities were created to focus on several key architecture issues.

Hitchcock discusses the normal operation of one system of automated freeways as described by Hsu. This system is being used as a basis for many system engineering studies within Partners for Advanced Transit and Highways (PATH). The system minimizes the degree to which the infrastructure is involved in minute-to-minute maneuvers. In it, each vehicle, as it enters, is given a route including lane choices to the destination. As described, however, no account is given of procedures on entry or exit or of possible faults. The PATH safety program demanded a second example of the process of full specification and fault tree analysis to determine if this process was generally applicable.

Rao et al. discuss the application of simulation SmartPath, which models the passage of individual vehicles along the highway. The simulator allows the researchers to examine transient behavior of the traffic stream under various conditions. Three different strategies are discussed for allowing vehicles to enter and leave automated lanes and measuring the maximum flow rates that are attained. The authors conclude that although maximum theoretical capacity cannot be attained, through prudent design of entrance and express strategies extremely high throughput can be sustained.

Rao and Varaiya examine the potential flow increases when only a proportion of vehicles on a highway are equipped with autonomous intelligent cruise control (AICC). The authors use a simulator that models interactions between vehicles to give detailed information on achievable capacity and traffic stream stability. The authors conclude that capacity gains from AICC are likely to be small.

Kaub and Rawls present a single and inexpensive IVHS speed monitoring concept that can be easily adapted to existing vehicles and roadways. The concept relies on the speed-distance-time relationship and on an on-board impulse detector and constant times to calculate the travel time or posted speed of the roadway.

Chira-Chavala and Lechner discuss the preliminary engineering feasibility of early deployment of a roadway-powered electric vehicle in El Monte Busway (a 3+ high-occupancy-vehicle facility in Los Angeles). The evaluation consists of determinations of the scale of electrification, the location to be electrified, the mode of operation, the level of energy transfer, and consumption of energy.

Chira-Chavala and Zhang discuss the phased implementation of advanced lateral guidance systems in high-occupancy vehicle lanes with exclusive right-of-way. Steering assistance information systems, partially automated lane-keeping systems, and fully automated lateral control systems are described.

Kaysi et al. propose a system structure for real-time traveler information systems consisting of a surveillance module, a congestion prediction module, and a control and routing module. The focus is on the approaches that may be used for congestion prediction and the strategies that may form the basis for routing.

Vaughn et al. discuss the result of an experiment to collect sequential route choice data under the influence of ATIS. The experiment collected information on drivers' pretrip route choice behavior at three levels of information accuracy: 60, 75, and 90 percent. The results of the experiment indicate that drivers can rapidly identify the accuracy level of information that is provided and that they adjust their behavior accordingly.

Mahmassani and Peeta present a comparative assessment of network cost and performance under time-dependent system optimal (SO) and user equilibrium (UE) assignment patterns, with reference to the effectiveness of ATIS. Both SO and UE solutions are found using a new simulation-based algorithm for the time-dependent assignment problem. The results of their work affirm the validity of a meaningful demarcation between system optimal and user equilibrium assignments in urban traffic networks and provide useful insights for macroscopic network-level relations among traffic descriptors.

Ziliaskopoulos and Mahmassani discuss an algorithm that calculates the time-dependent shortest paths from all modes in a network to a given destination mode for every time step over a given time horizon in a network with time-dependent arc costs. The motivation for this study was the need to compute time-dependent shortest paths in a real-time environment in connection with IVHS. The suitability of the proposed algorithm for such applications is demonstrated in this study.

Chadwick et al. propose a communication architecture for IVHS called the subsidiary communications authority traffic information channel (STIC) on the basis of the widely available FM radio broadcast services infrastructure by making use of FM subcarrier technology. According to the authors, the STIC has a higher data transmission capacity than any other existing FM subcarrier broadcast system, and it has the potential to meet the one-way outbound (broadcast) data transmission capacity needs of IVHS for the next few years.

Michalopoulos et al. indicate that machine vision is one of the most promising technologies for developing and deploying ATMS and ATIS applications. The paper describes the first and largest application of video detection and its integration with adaptive control in a network of 28 intersections. The broader project (FAST-TRAC) is also summarized.

Bullock and Hendrickson discuss why improved traffic controllers will be essential for many proposed IVHS applications. The authors introduce a computer language called Traffic Control Blocks (TCBLKS) that could provide the foundation for constructing real-time traffic engineering software.

Intelligent Vehicle Highway System Benefits Assessment Framework

DANIEL BRAND

A framework of linked cause and effect relationships (models) is derived for use in intelligent vehicle highway systems (IVHS) project and operational test evaluation. IVHS has the potential for greatly increased mobility, measured in travel opportunities and benefits, as well as considerable potential for improved transportation system operation. The framework avoids serious underestimation of the mobility and other user benefits from IVHS and allows the estimation of the effects of IVHS on aggregate volumes of travel and levels of congestion. The mobility benefits of IVHS must be measured at the level of the individual tripmaker, not on the basis of aggregate measures of flow volumes and travel times on the network. This means that the air pollution, safety, fuel consumption, and other flow volume–related impacts of IVHS do not vary in a straightforward way with the sum of the individual user benefits from IVHS. Therefore, the causal model chain for predicting IVHS impacts will vary from the conventional "planning model." The various predictive models required to evaluate IVHS improvements, including the formulation of model inputs, are described. The evaluation framework is intended to help guide the evaluation and selection of IVHS projects on the basis of their site-specific benefits and costs, rather than the desired results. Although the latter is entirely acceptable for planning a research program whose payoff cannot be known in advance, it is necessary to proceed to the next step of carefully evaluating operational field tests and advancing IVHS into its production mode. The causal framework makes it possible to anticipate the important consequences of IVHS and therefore carry out benefit-cost analyses of new investments as well as collect the appropriate data for planning and evaluating operational field tests.

This paper derives a framework of cause and effect relationships (models) for use in intelligent vehicle highway systems (IVHS) project and operational test evaluation. Both of these require appropriate causal models explaining IVHS impacts. If the important consequences of IVHS cannot be anticipated properly, benefit-cost analyses of new investments cannot be carried out, and the appropriate data for evaluating operational field tests cannot be collected.

IVHS differs considerably from conventional transportation capacity increases and operational improvements. What differentiates IVHS strategies from conventional transportation improvements is the development of a user-friendly information infrastructure to complement and increase the productivity of our massive investment in transportation infrastructure *(1)*. The new user and information orientation in transportation, together with the rapid pace of technological change, accounts for much of the current excitement in transportation and the dramatic increase in the number of transportation options being considered today *(2)*. Les Lamm, the executive director of the Highway Users Federation and president of IVHS America, stated at the 1991 TRB Annual Meeting that "IVHS is the most significant transportation initiative of my generation." This puts IVHS in the same category as the Interstate highway system in the promise and excitement it holds for transportation in America.

One would imagine that the benefits of an important transportation program such as IVHS to travelers and society would be well known by now. The truth is that people may be as ignorant now of the consequences of IVHS as they were of the impacts of the Interstate highway program at its inception in the 1950s. The IVHS American "Benefits, Evaluation and Costs" Committee is on record that "substantially lacking are defensible methods for predicting changes (in benefits and costs) which could be brought about as a result of IVHS technology deployment (B. Stephens, personal communication to J. Vostrez, Dec. 2, 1991).

However, transportation planning has changed considerably since the 1950s. Planners now strive to carry out rational investment planning using benefit-cost techniques in an analytic framework. Changed also are the statutory requirements under which transportation improvements are made. The 1990 Clean Air Act Amendments (CAAA) require explicit consideration of whether transportation improvements produce more, rather than less, air pollution. Indeed, the CAAA established the principle of regional emission budgets and conformity to the emission reduction schedules contained in state implementation plans. With the exception of the 1987 congressionally mandated cost-effectiveness requirement for major transit investments, this appears to be the first significant, federally imposed regulatory performance standard for new transportation investments.

The importance of clean air and IVHS, as well as the uncertainty surrounding the impacts of IVHS as a transportation investment, suggests that improved methods for assessing the benefits of IVHS are required. The key impacts of IVHS must be brought together in a framework that will allow credible estimates of IVHS costs and benefits to be developed. However, the air pollution, fuel consumption, safety, and other travel volume–related impacts of IVHS do not vary in a straightforward way with the sum of the individual user benefits from IVHS. Increased user benefits from travel usually lead to more travel, which then gives rise to more impacts of this travel. Because this direct relationship is not the case with IVHS, as will be shown, the causal framework will differ from the conventional planning model shown in Figure 1.

Charles River Associates, 200 Clarendon Street, Boston, Mass. 02116-5092.

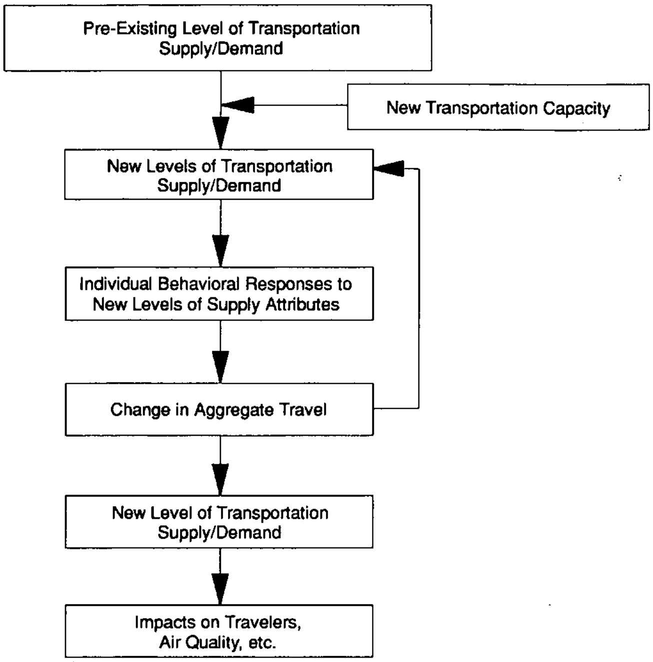


FIGURE 1 Conventional planning model for assessing benefits from changes in transportation capacity.

RELATIONSHIP OF TRANSPORTATION BENEFITS TO IVHS INFORMATION AND CONTROL

IVHS differs from conventional transportation improvements in the way information is communicated and used to increase the benefits from travel and system operation. Information is communicated in real time to the traveler on the transportation system status and operation and on travel services and trip end opportunities. Information is also communicated in real time to the system to improve its traffic (and ultimately its vehicle) control capabilities. Figure 2 shows the relationship of IVHS transportation benefits to various levels of information and control. Control—which is defined here as mandatory—is applicable only at signalized intersections and metered ramps and with degrees of roadway (or guideway) automation. Only vehicles are subject to mandatory control (in a democratic system).

Information/ Control	Applicable		Benefits
	To	Where	
Control	Vehicles	Intersections and metered ramps	Street and freeway link capacity maximization
Information	Travelers	Pretrip or en route	Multimodal network flow optimization
Control	Vehicles	En route	Link and network capacity maximization through automatic control

FIGURE 2 Relationship of transportation benefits to IVHS information and control (I).

On the other hand, as shown in Figure 2, information on highway and multimodal network status, travel conditions, routing, and guidance applies to people (not vehicles). Individual travelers use this information to plan and make all of their travel choices (e.g., mode, time of day, destination, and whether to make the trip at all)—not only the path or mode choice for which the information may be available.

Figure 2 shows that the transportation level-of-service and capacity benefits increase with improved information to the traveler and greater control of the vehicles. Fully automated vehicle operation (with resulting increases in information) will maximize link and network capacity. With automatic vehicle control, it is assumed that users will give up their route and mode control in a pretrip choice to use the system because it provides significant travel time, safety, and other benefits through guideway automation. If users trust the system, they will accept the guidance information on their multimodal options. If they trust the system even more, they will entrust their lives to an automated guideway that maximizes throughput.

RELATIONSHIP OF USER BENEFITS TO IVHS INFORMATION AND CONTROL

The inevitable consequence of the IVHS information infrastructure that will parallel the transportation infrastructure is a paradigm shift in how mobility is measured. When the sole concern was improving the physical transportation infrastructure, improvements were evaluated on the basis of the use of the network. Vehicle miles traveled (VMT) on the network and congestion and travel times on links were the measures of interest. With the development of a parallel information infrastructure, parallel emphasis must be on the use of the information, that is, how travelers use the information to make their travel decisions. Mobility, which is what travelers seek, is measured by the opportunities for, and the benefits from, travel. Figure 3 adds this dimension of information and mobility benefits to the more limited set of network flow benefits shown in Figure 2.

Information/ Control	Applicable		Benefits
	To	Where	
Control	Vehicles	Intersections and metered ramps	Street and freeway link capacity maximization
Information and conditions: Travel modes	Travelers	Pretrip or en route	Multimodal network flow optimization
Trip-end activities/ services	Travelers	Pretrip or en route	Mobility maximization
Control	Vehicles	En route	Link and network capacity maximization through automatic control

FIGURE 3 Relationship of total benefits to IVHS information and control (II).

With IVHS, plans can be made to maximize the mobility benefits from travel, rather than produce an elusive level-of-service performance standard. One can anticipate that the new transportation information environment created through IVHS will provide more benefits as a result of the information it gives to travelers than from the shortened trip times it provides. This will result in higher-value use of personal time and resources for work and leisure activities and more productive use of commercial and industrial resources.

This means that IVHS systems will reduce or modify demand in response to the information they provide on congestion and trip end opportunities while providing benefits from increases in effective network capacity. Their contributions to user benefits will therefore result from demand management at the level of trip generation, as well as at the level of route and mode choice, which makes more effective use of existing capacity in the network. In other words, the benefits from such systems will come as much or more from user interactions with the system as from increases in effective network capacity that these systems supply. How this happens is discussed in the next section.

IVHS BENEFITS ASSESSMENT FRAMEWORK

Travel decisions involve a series of trade-offs between the times and costs of travel on all available alternatives and the benefits of travel from engaging in activities at the trip ends. Without adding capacity, the information from IVHS will increase the informed nature of these trade-offs and all of the adjustments people make to minimize their cost of travel (e.g., avoiding congestion). For example, with reliable travel time information, many travelers for whom the benefits of certain trips are small will choose to travel shorter distances, change modes, or forego or defer trips when congestion is heavy. Others may choose to travel to destinations that are farther away or else make more frequent trips with the confidence that they will not be caught in heavy congestion. The net increase in user travel benefits from these travel decisions will be substantial, yet aggregate reductions in VMT and travel times are not likely to reflect these benefits. In fact, the aggregate reductions are likely to be small. They may even be negative.

Another type of IVHS system is one that provides reliable attraction location information and travel directions to unfamiliar drivers at popular tourist destinations (for example, as supplied by the 1992–1993 Travtek demonstration in Orlando). This system is intended to improve travel routing efficiency and minimize the time one spends lost in a strange city. The system, however, is also likely to encourage tourists to visit more attractions and increase the entertainment value of tourists' vacations. Aggregate VMT and time spent traveling may increase, but mobility and user benefits will increase even more. It is reasonable to conclude that in this case also, the user benefits of IVHS will be much greater than those resulting from reductions (if any) in aggregate travel time and delay.

The traveler utility functions used in assessing the benefits from, and forecasting the travel impacts of, IVHS must therefore be redefined. Doing so will avoid serious underestimation of the user benefits of IVHS while allowing accurate estimation of the resulting aggregate numbers of trips and the link-specific vehicle volumes and conditions of travel on which the flow-related physical impacts of IVHS are based (e.g., accidents, fuel use, air quality). These concepts are brought together in Figure 4, which shows the proposed IVHS benefits assessment framework.

Differences from Conventional Planning Model

The major difference between the IVHS benefits assessment framework in Figure 4 and the conventional "planning model" in Figure 1 is that the travel volumes and flow conditions used in calculating the impacts on air quality, energy use, and so on, of an IVHS improvement are not used in calculating the user benefits from the improvement. The IVHS evaluation framework splits the benefit calculations in two parts, with the introduction of real-time travel and trip-end opportunity information provided by IVHS. User benefits are calculated separately from the link travel volumes in the behavioral travel demand model. Therefore, the air pollution, fuel consumption, and other travel volume–related impacts of IVHS are not directly related to the sum of the individual user benefits of IVHS. (Note, however, that Figures 1 and 4 both adhere to the same supply and demand mechanism governing the assessment of benefits from transportation improvements.)

IVHS evaluation therefore requires that changes be made to the data input to the travel demand models. Current travel models are based on average travel times and costs on highway links or scheduled times on transit links. Travel forecasts to evaluate IVHS systems must receive, as input from a traffic simulation model, the travel times and costs on the various

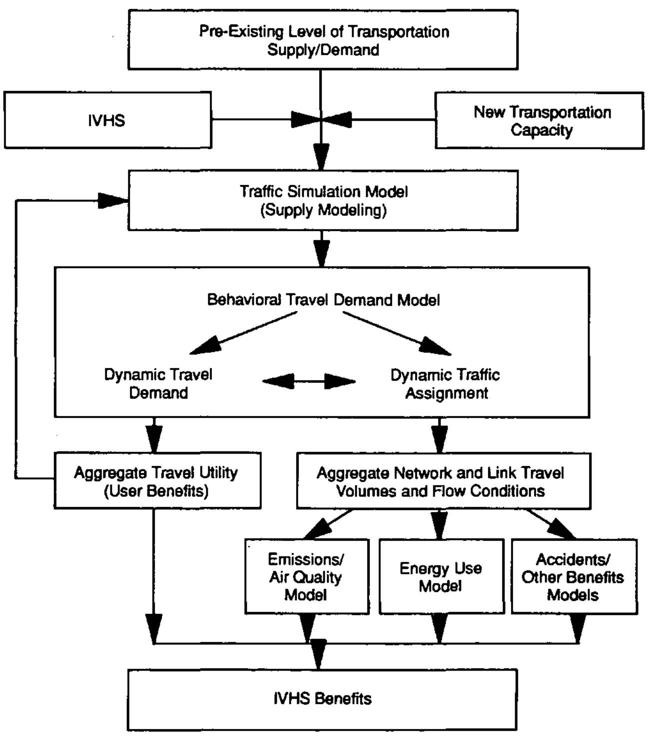


FIGURE 4 IVHS benefits assessment framework.

travel choices provided to travelers in real time. This avoids serious underestimation of the user benefits of IVHS while allowing estimation of the resulting aggregate numbers of trips and the link-specific vehicle volumes and conditions of travel on which the flow-produced physical impacts (e.g., fuel use, safety, and air quality) of IVHS are based.

Demand models for evaluating IVHS must also take into account the fact that information given to travelers on, for instance, highway congestion and route guidance may affect all the travel choices of travelers, not only the route choice for which the information is provided. Behavioral travel models must be able to forecast these changes in trip frequency, destination choice, departure time, mode choice, and so forth. These changes must be predicted at the level of the behavioral unit—the trip itself. Similarly, when evaluating an operational test, the model cannot be restricted to anticipating and measuring only changes in path characteristics (e.g., link volume and travel times). To do so would be to overlook many of the important impacts of IVHS improvements.

Specification of Traveler Utility Functions

The utility functions for our travel demand models must also incorporate variables representing the mobility benefits from engaging in the trip-end activities in the traveler's newly expanded choice set. These variables are required to calculate the net increase in benefits from, for example, high-value trips to destinations further away, or trips made more frequently with the information from IVHS that the traveler will not be caught in heavy congestion. If motility benefits are not included, comparisons between the new and old (without IVHS) travel times may show negative user benefits from the longer (higher-value) trips made with improved travel information from IVHS.

The utility functions for demand models also will need to incorporate not only the "traditional" travel time and cost variables but also variables describing the IVHS information itself. These may include human factor variables on information sequencing and display and variables describing the new dimensions that the information adds to the value of the traditional level-of-service variables normally included in demand-model utility functions.

The two most important examples of the new variables are likely to be the reliability of the information and the degree of control over the person's time and life that the new information provides. Travel time reliability has been recognized for decades as an important variable for travel demand forecasting. Its valuation (quantification) in traveler utility functions, however, has been almost entirely lacking. It is currently measured by the variance in travel times on a highway or transit (door-to-door) route. The traveler perceives reliability as the difference between his or her average travel time and day-to-day travel time actually experienced.

There is considerable psychological evidence that a few large negative travel time experiences are much more highly valued than many positive values (negative reinforcement). This is consistent with evidence that as they come to depend on more reliable information, travelers value it highly. This has already happened in logistics and with overnight package delivery. Small package delivery companies track package movements at every step. The fax machine is another example of how new technology has escalated our concern for service quality.

In the case of IVHS, the transportation system will be more able to respond in its allocations of highway and transit capacity to additional traffic and passenger loadings, thus decreasing the severity of travel time fluctuations under many conditions. In addition, as IVHS systems monitor travel conditions in real time and are improved to the extent that they can predict future system loadings and future travel conditions, they will significantly reduce the error in the travel times on the various travel choices presented to travelers by IVHS (i.e., the variance between the travel time presented to travelers and the times they actually encounter).

As important as time is, the most important variable to be considered may be the added control over time that IVHS may ultimately provide. It has been stated that "metropolitan areas have a strong hold on the externalities that promote population growth. Suburbanites want control over the temporal and spatial dimensions of their travel and will pay large sums of money for these" *(3)*.

The desire to control one's use of time will increase as one's ability to control it increases. IVHS will allow travelers to control, or at least better manage, their use of time at their trip destinations and the levels of congestion or delays that characterize their travel to those destinations. Thus the control over time that IVHS will provide should make travel time even more valuable.

Valuing IVHS User Benefits

Valuing the user benefits from IVHS requires the use of behavioral travel demand models, as shown in Figure 4. As noted earlier, demand models and survey data at the level of the individual traveler are needed to measure the mobility improvements provided by IVHS. The disaggregate data required to estimate the demand models can be developed from stated preference surveys and carefully evaluated operational tests.

In general, the only transportation system change that should be included in valuing the user benefit of a transportation improvement is the item being changed. For example, in an evaluation of a conventional transit improvement, the only transportation system change that should be input to the demand model is the transit system change. User benefits from this change accrue both to existing transit users and to automobile users who divert to transit as a result of the change. However, the user benefit from an automobile trip diverted to transit is only the change in the utility function value used by the traveler in deciding to switch from automobile to transit. The automobile costs do not enter into this change; only the transit "costs" change. This applies even if automobile service is improved in the process.

In the IVHS context, the change is the (valid) information on the chosen trip, including the trip-end benefit from the trip. A person makes the travel choice that yields the highest utility from weighting the variables in his or her utility function. The utility function is used in the behavioral model to calculate the change in traveler benefit, which is needed to explain (cause) the shift in travel behavior (e.g., route, mode, departure time) predicted for the IVHS improvement or observed in the IVHS operational test. Valuing the user benefit includes the user's valuation of change in all of the variables included in the properly specified traveler utility function. The change can be valued monetarily at the change in travel cost (price) that is equivalent to the change in the traveler's utility. Half of this value can be used in the usual manner to quantify the area in the consumer surplus triangle relating to induced travel, whereas the entire area in the consumer surplus rectangle can be used to calculate the value of the IVHS benefits for existing travel.

These observations are made here to emphasize that a proper accounting of the user benefits of IVHS improvements is very important. The changes being evaluated are the IVHS information presented to the traveler and the improvements in system operation made possible with this information. Valuation of these benefits from the IVHS improvement must include the changes in all of the mobility and information attributes in properly specified traveler utility functions that are incorporated in behavioral travel demand models. Significant improvements in these models are needed, as discussed in the next section.

Behavioral Travel Demand Models

In the evaluation framework shown in Figure 4, the behavioral demand model is used to forecast the changes in all of the travel choices (path, mode, departure time, trip frequency, destination) to

- Quantify the user benefits from the information supplied to travelers and the improved system performance and
- Calculate the flow volumes and travel conditions on links in the network, which are required as inputs to the models of flow-related physical impacts.

Behavioral travel demand models are needed that explain individual travel behavior. Figure 5 illustrates a model of individual behavior that incorporates IVHS information on activity and travel opportunities *(4)*. The individual has information about a set of opportunities to engage in activities at various locations, some or all of which may involve travel.

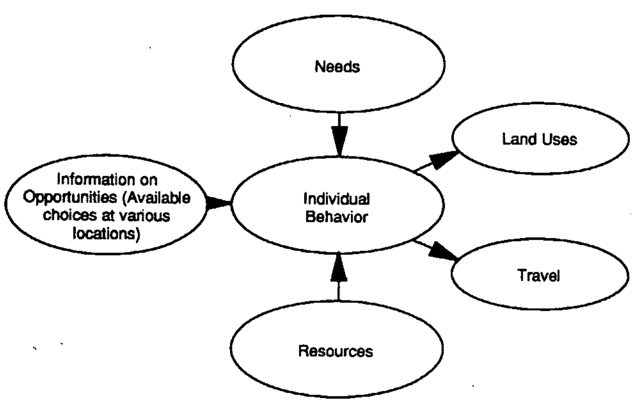


FIGURE 5 Suggested paradigm of individual behavior incorporating IVHS information.

The individual also has needs—to work, shop, play, and be safe and also to have a home. These condition how the individual chooses from among various activity opportunities that involve travel. The individual also has resources (e.g., time and money) that affect his or her response to opportunities to travel and engage in activities at various places and prices.

The lack of a direct causal relationship between land use and travel is shown in Figure 5. A third variable drives them both, namely individuals responding to opportunities, needs, and resources to "consume" both land and travel. Empirically, the presence of the third variable has been amply demonstrated; individuals consume both more land and more travel as their income increases *(5)*.

The sequential aggregate travel demand models used today in urban transportation planning are well known to be highly deficient in their sensitivity to changes in even conventional transportation capacity. Trip generation equations are almost always totally insensitive to travel conditions. Trip distribution is modeled as a function of a simple description of trip lengths that prevailed at the equilibrium between supply and demand, represented in trip data, and so on *(6)*. These sequential models are not adequate to evaluate IVHS improvements, both because they are not capable of incorporating properly specified travel utility functions and because they are much too cumbersome to operate in the context of dynamically changing travel conditions.

More likely than not, the selected models of individual travel choice behavior incorporating the traveler utility functions that are necessary for valuing user benefits will be "direct demand" models. Current direct (travel) demand models forecast travel directly by mode between origins and destinations as a function of the activity systems at the origins and destinations, and the price and level-of-service conditions by the travel mode and all its substitutes *(7)*. These direct demand models are themselves simplifications of general equilibrium models that explain how land use and travel vary simultaneously with transportation improvements (8). They are partial equilibrium models that describe how part of the system behaves so it will be in equilibrium with the rest of the system. Thus, there is modeling of the behavior of the tripmaker, who considers all trip end opportunities to be fixed. This may or may not be appropriate for IVHS evaluation, depending on how the models' relationships between travel and its determinants are structured. The paradigm in Figure 5 suggests that the evaluating IVHS systems, developing models that incorporate real-time information on dynamically changing travel opportunities and costs may be at least as important as developing general equilibrium models.

Dynamic Travel Models

The IVHS benefits assessment framework in Figure 4 shows that traffic simulation models are used to input initial "supply" conditions into the behavioral travel demand model. For example, it is relatively straightforward to model the effects of improved traffic signal settings from IVHS (ATMS) on a fixed set of link-traffic volumes, using currently available tools such as NETSIM and TRAF-NETSIM. However, because queuing is so characteristic of congested highways, assuming link flows and travel times to be time invariant does not accurately describe the stochastic nature of congested traffic. For this reason, the framework in Figure 4 separates the initial simulation models from the later dynamic travel models.

Dynamic travel models are intended to analyze the effects of IVHS information on travel volumes and congestion levels in high-volume networks. The state of the art in dynamic assignment and simulation for IVHS is summarized by Mahmasani et al. *(9)*. At present, most dynamic travel models are highway path-choice models, which provide the ability to

- Model the route-choice behavior of drivers with and without access to IVHS information;
- Predict travel times on the basis of the assignment results and provide feedback to the control center that may be used in the assignment of vehicles; and
- Track the location of the drivers who receive guidance information from the control center.

Work on dynamic traffic assignment is moving rapidly for descriptive user equilibrium and normative system optimizing problems. Key research areas are modeling and incorporating the appropriate travel behavior decisions and representing the dynamic (transient) traffic phenomena of congested networks.

Ultimately, dynamic traffic assignment (path choice) models will be fully integrated into the behavioral demand model, as shown in Figure 4. For an example of a model system that combines mode, departure time, and route choice in a dynamic model, with interdependent travel costs, see Boyce et al. *(10)*. The interaction of these models with the models of the other travel choices will be complex because the objective is to model (explain) not only path choice but also, as discussed earlier, the behavior of individuals making, for example, high-value trips to destinations farther away, or trips made more frequently with the information from IVHS that the traveler will not be caught in heavy congestion. The behavioral response to this information is anything but fixed. To be able to model (or simulate) it in a time-varying (dynamic) context, it is necessary to understand the cause-and-effect mechanisms that govern travel behavior. Unfortunately, empirical data on these behaviors are as yet extremely limited.

Regardless of whether they describe time variant or invariant travel and flow conditions, the behavioral travel models used for IVHS evaluation must contain appropriate supply and demand equilibration mechanisms that ensure that the information given to travelers (and input to the travel model) is the same as that produced by the model. This is the only way to assess whether IVHS can permit travel at more efficient speeds and thereby reduce air pollution, and so on. The linkages between all of the travel choice models, however dynamic, must converge on a steady-state output, albeit with the required stochastic distributions of traffic characteristics. This does not preclude modeling the response of travelers over time in all of the travel choice models (including dynamic traffic assignment), as they receive information on time variant travel and trip-end activity conditions.

FLOW-PRODUCED PHYSICAL IMPACT MODELS

The final set of models in the evaluation framework in Figure 4 are the models linking the emissions, safety, fuel use, noise, and other flow-produced physical impacts with the travel volumes and flow characteristics resulting from the IVHS improvement. For example, to evaluate whether candidate IVHS strategies are compatible with CAAA requirements, the framework links the travel demand model with the mobile source emissions models to assess air quality impacts. Currently available emissions and other impact models that use flow volumes and travel conditions as inputs are relatively easy to adapt to an IVHS evaluation.

Measuring the mobility improvements from IVHS on the basis of the behavioral unit of travel—namely the trip—has advantages not only for evaluating user benefits but also for measuring air quality impacts. For example, well over half of vehicle emissions from even a long (20-mi) summer automobile commute are caused by the trip being made in the first place—the combination of trip start emissions and hot soak emissions at the trip end—not by the VMT on the links of the network.

Ultimately, it is important to recognize that estimation of carbon monoxide and hydrocarbon emissions (for example) from motor vehicles is a complex process that requires a substantial amount of information on the amount and type of travel, vehicle activity (start, cruise, idle, acceleration, deceleration, time between starts and trips), vehicle characteristics (type of vehicle, age, size, engine type, transmission system, antipollution devices, type of gasoline used), vehicle operating conditions (hot versus cold starts, engine temperature, vehicle load, speed of vehicle), environmental conditions (altitude, ambient temperature), and roadway conditions (horizontal and vertical alignments). An accurate estimation of the pollutants from vehicle emissions may necessitate analysis at a very detailed level, beyond the resources and data available for typical transportation planning studies.

Institutional factors will also dictate the inclusion of certain variables and relationships in the framework. These range from required output variables (e.g., specifically mandated air quality measures) to restrictions on the way the system can operate (e.g., liability governing when and how IVHS information can be presented) to whether IVHS develops as a stand-alone system or becomes part of a metropolitan information utility with consumers trading off many activities that compete for their time and money (potentially involving new layers of variables and relationships).

IMPLEMENTING IVHS EVALUATIONS

The evaluation modeling framework presented in this paper is designed explicitly to quantify the benefits of IVHS, including additional benefit measures to those from conventional transportation improvements, to avoid seriously underestimating the benefits from these improvements. Figure 6 presents a list of evaluation measures relevant for evaluating IVHS improvements that are in addition to those normally used to evaluate conventional transportation improvements. The additional user benefit measures in Figure 6 are the primary focus of this paper.

Ultimately, what is included in IVHS evaluations is linked to the goals set for IVHS: what travelers want and expect IVHS to accomplish. For example, some proponents of IVHS see the potential of IVHS strategies for internalizing some of the current external (social) costs of congested highway travel.

For Conventional Transportation Improvements:

- User Benefits
  - Travel time
  - Travel cost
  - Safety
- Costs
  - Construction (capital)
  - O&M
- Externalities
  - Air quality
  - Fuel/energy consumption
  - Noise
  - Land use patterns
  - Neighborhood impacts
  - Productivity/economic development

For IVHS Add:

- User Benefits
  - Benefits from trip end opportunities
  - Travel time reliability
  - Control
  - Privacy
  - Legal
- Implementation Risk
  - Feasibility
  - Technology flexibility
  - Ease of implementation and use (staffing/skills)
  - Community acceptance
  - Agency Coop./Coord.
- Other Benefits
  - Improved data collection

FIGURE 6 Additional measures needed to evaluate IVHS improvements.

Individuals currently perceive only a fraction of the total congestion they cause. Every time a driver enters a heavily congested roadway, far more aggregate delay is imposed on others—on the system—than on that driver. In turn, this aggregate delay results in far more air pollution and energy consumption by others than by the individual causing the delay and pollution in the first place. In fact, the more congested the highway, the greater the difference between the social and private costs of making an additional or longer trip by automobile *(12)*.

Congestion is also the price that the current transportation system imposes on everyone as a result of individual life-style decisions to locate in sprawling regions and on larger plots of land, farther away from work and shopping. And because increasing amounts of money are spent on housing, the transportation price that individual life-style decisions impose on everyone else is not known by the individual making those decisions. Individuals make investments in expensive housing without considering the total cost of their location decisions *(4)*. IVHS is likely to play a role in promoting more informed activity location decisions in the long run, just as it informs such decisions in the short run.

The evaluation framework presented in this paper is intended to help guide the evaluation and selection of IVHS projects on the basis of their site-specific benefits and costs, rather than their desired results. Although the latter is entirely acceptable for planning a research program whose payoff cannot be known in advance, it is necessary to proceed to the next step of carefully evaluating operational field tests and advancing IVHS into its production mode. The causal framework described in this paper allows one to anticipate the important consequences of IVHS and therefore carry out benefit-cost analyses of new investments as well as collect the appropriate data for planning and evaluating operational field tests.

REFERENCES

1. *Special Report 232: Advanced Vehicle and Highway Technologies*. TRB, National Research Council, Washington, D.C., 1991.
2. D. Brand. Conference Summary. In *Special Report 237: Moving Urban America*, TRB, National Research Council, Washington, D.C. 1993.
3. J. D. Kasarda. Population and Employment Change in the United States: Past, Present and Future. In *Special Report 220: A Look Ahead: Year 2020*, TRB, National Research Council, Washington, D.C., 1988, pp. 83–148.
4. D. Brand. Research Needs for Analyzing the Impacts of Transportation Options on Urban Form and the Environment. In *Special Report 231: Transportation, Urban Form, and the Environment*, TRB, National Research Council, Washington, D.C., 1991, pp. 103–107.
5. E. A. Deakin. Jobs, Housing and Transportation: Theory and Evidence on Interactions Between Land Use and Transportation. In *Special Report 231: Transportation, Urban Form, and the Environment*, TRB, National Research Council, Washington, D.C., 1991, pp. 25–42.
6. D. Brand. Use of Travel Forecasting Models To Evaluate the Travel and Environmental Effects of Added Transportation Capacity. Presented at DOT/EPA Conference on the Travel and Environmental Effects of Added Transportation Capacity, Bethesda, Md., 1991.
7. T. A. Domencich, G. Kraft, and J.-P. Valette. Estimation of Urban Passenger Travel Behavior: An Economic Demand Model. In *Highway Research Record 238*, HRB, National Research Council, Washington, D.C., 1968.
8. D. Brand. Theory and Method in Land Use and Travel Forecasting. In *Transportation Research Record 422*, TRB, National Research Council, Washington, D.C., 1973.
9. H. Mahmasani et al. *A Review of Dynamic Assignment and Traffic Simulation Models for ATIS/ATMS Application*, The University of Texas at Austin, 1992.
10. D. Boyce, B. Ran, and L. LeBlanc. *A Combined Dynamic Model Departure/Route Choice Model with Multiple Groups of Travelers*. Presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1993.
11. P. De Corla-Souza. *Surburban/Urban Transportation Investment—What Will It Cost and Who Will Pay*. FHWA, U.S. Department of Transportation, 1990.

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Comparison of Advanced Traffic Management and Traveler Information System Architectures for Intelligent Vehicle Highway Systems

MELVYN D. CHESLOW AND S. GREGORY HATCHER

One of the important efforts in the intelligent vehicle highway systems (IVHS) program is the development of a system architecture to guide development and implementation decisions. Five alternative architectures for advanced traffic management and advanced traveler information systems are compared. These architectures were created to focus on several key architectural issues. The qualitative evaluation focuses on characteristics of the alternative architectures that affect performance, cost, and risk. Additionally, the evaluation stresses issues connected with starting up IVHS services and evolving to more advanced systems and addresses institutional concerns.

The U.S. Department of Transportation through FHWA has established the intelligent vehicle highway systems (IVHS) program to use advanced technology to improve the efficiency of the nation's highway transportation resources (*1*). IVHS can be defined as advanced communications, navigation, sensors, control systems, and information systems that can be used to increase throughput on existing roadways, to improve the safety of the traveling public, and to improve the productivity of commercial vehicle operations. The IVHS program is divided into six major areas:

- Advanced traffic management systems (ATMS),
- Advanced traveler information systems (ATIS),
- Advanced vehicle control systems (AVCS),
- Advanced public transportation systems (APTS),
- Commercial vehicle operations (CVO), and
- Advanced rural transportation systems.

One of the early requirements in the IVHS program is a system architecture to guide development and implementation decisions. The architecture provides the framework, or structure, for defining the functions that provide IVHS services. It also defines the information flows and interfaces between functions. The system architecture, in addition, describes the way the IVHS functions are divided among subsystems in the vehicle, at the roadside, or in one or more traffic management centers (TMCs).

A process for defining and implementing an IVHS architecture is shown in Figure 1. The basic steps are to

- Define goals and objectives for IVHS systems,
- Identify alternative functions that meet these objectives,
- Synthesize alternative architectures that may provide these functions,
- Assess and refine the alternatives,
- Select an acceptable architecture, and
- Design and implement the systems in accordance with the architecture.

Numerous organizations have developed definitions of IVHS goals and identified candidate solutions (*1*). Yablonski has provided an extensive discussion of IVHS services and functions (*2*). Several alternative architectures have been documented in earlier work (*3*). The assessment and refinement process shown in Figure 1 must be carried out in an iterative fashion and will require several iterations to come to an acceptable architecture.

This paper describes initial results for Steps 3 and 4 of the architecture development process: the synthesis and assessment of five alternative IVHS architectures. The alternative architectures are partial architectures because they have been limited to providing only major ATMS and ATIS functions and do not include the provision of AVCS, APTS, or CVO services. They can be described as end-state architectures because they have been defined to support the provision of fully developed ATMS and ATIS services. The alternatives have been named strawman architectures for two reasons. The first is that they were created to focus initially on several key architectural issues, rather than encompass all the ATMS and ATIS functions and services. The second is that the evaluation of these initial architectures will lead to modified and improved architectures.

There is some question about how to actually evaluate system architectures because they are not fully defined designs. They only provide structures for implementing the required services. Hence, it is not clear that one can, or should, carry out a traditional analysis of the benefits and costs for an architecture. It was therefore decided to carry out a qualitative comparison of several important characteristics of the architectures that would directly affect their eventual performance and costs. This evaluation is discussed more fully in a report by Cheslow et al. (*4*).

With current technical and institutional uncertainty, it is not possible to state specifically that one future end-state architecture is best. Instead, the evaluation described here

The MITRE Corporation, 600 Maryland Ave., S.W., Suite 755, Washington, D.C. 20024.

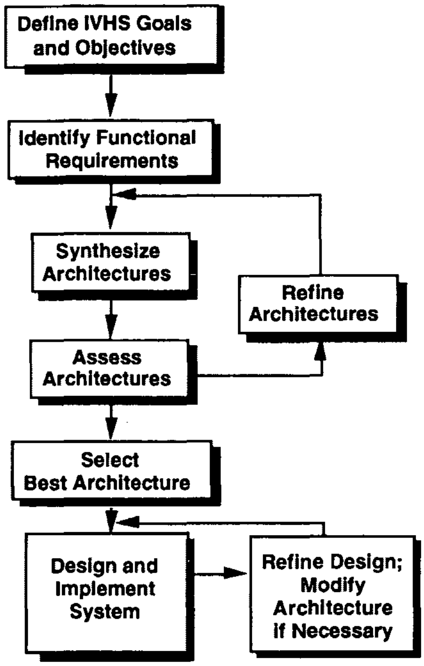


FIGURE 1 IVHS architecture selection process.

provides some initial conclusions about which types of end-state architectures are promising. Further analyses of IVHS architectures, as well as related research and development and operational tests, will eventually provide the necessary quantitative information to allow the determination of the best architecture.

A summary of the five strawman architectures is presented next, beginning with a description of the key architectural attributes that define them. A complete discussion of these alternatives, along with the approach used to develop them, is documented in earlier work (*3*). These five architectures highlight various functional approaches to providing ATMS and ATIS services, ranging from highly centralized to fully distributed. The architecture that finally evolves may not be identical to any of the original strawman architectures; it may combine features of two or more of the alternatives or may add features not found in any of the five.

KEY ATTRIBUTES OF THE STRAWMAN ARCHITECTURES

The architecture definition process focused on three critical system elements that highlight major differences among the alternatives. These three elements are

- Vehicle-infrastructure communication alternatives,
- Location of the route selection function, and
- Degree of coupling between route selection and traffic control.

Although a single vehicle-infrastructure communication alternative is associated with each strawman architecture, the applicability of each of the communication alternatives to all of the architectures will be addressed in the paper. Further discussion of architectural issues can be found in other papers (*3, 5*).

Vehicle-Infrastructure Communications

A critical implementation issue for the IVHS program is the choice of an effective and economical approach for communicating traffic information and safety advisories to vehicles and collecting traffic flow information and assistance requests from vehicles. The strawman architectures have used four communication alternatives:

- Two-way localized beacons,
- Two-way wide-area coverage radio systems,
- Two-way cellular-like radio systems, and
- One-way broadcast radio systems utilizing an FM subcarrier.

These four options were chosen to represent a range of technological capability. Each approach has its specific advantages and disadvantages for supporting IVHS functions and services. With one exception, the communication options that were selected represent alternative ways of providing two-way communication services. Two-way communications between the traffic management infrastructure and vehicles will be required to enable the transmission of vehicle probe reports and assistance requests by vehicles and to receive traffic information and either traffic network link times or recommended routes from the infrastructure. Satellite-based communications were not considered for the strawman ATIS/ATMS architectures. In addition, alternatives for communication among the infrastructure components have not been considered. The following paragraphs summarize each of the four communication alternatives.

1. Localized beacons on the roadside support short-range, two-way transfer of traffic and routing information to or from vehicles that are near the beacon. The term beacon is used generically in this paper to mean any localized two-way communication device. Communication between the vehicle and the infrastructure can occur only at beacon sites. A variety of communication wavelengths, including infrared, radio, and microwave, could be used for the actual beacon implementation. Localized beacons have been chosen as a communication option because of their potential for spatially distributing the communication loads on the infrastructure. This approach can permit more location-specific information to be presented to vehicles within range of the beacon and would allow a large number of vehicles to be handled with a relatively small use of the electromagnetic spectrum.

2. A wide-area radio broadcast system is another alternative for two-way communications between vehicles and the infrastructure. The communication range is sufficiently large to cover a substantial portion of a metropolitan area (or critical segments of Interstate highways or rural areas). One or more dedicated pairs of radio frequencies is used to transmit and receive traffic information. This option represents a centralized communication system approach to providing two-

way communication coverage to or from any location within the area of interest. A wide-area radio system would broadcast the same traffic information to all vehicles within receiving range, relying on in-vehicle processors to separate out the relevant data on the basis of the vehicle location and planned route of travel. Each vehicle would also send (broadcast) information to the TMC.

3. A digital cellular radio system also supports two-way communications between vehicles and the infrastructure. The systemwide communication coverage is sufficiently large to serve a metropolitan area. However, the coverage area is broken into smaller cells, each served by a base station and a transmitter of lower power than that used with the wide-area radio option. This system provides more communication capacity for a given allocation of frequencies than does wide-area radio. A set of communication channels is reused many times throughout the network. This option represents a more distributed communication architecture than one based on wide-area radio.

Although analog cellular phone systems still dominate the market today, digital cellular systems will soon be common. Digital technology is assumed for this paper because of the significant capacity advantage it offers over analog. Two implementations of the cellular system infrastructure and spectrum are possible: (a) fully shared with conventional cellular applications (e.g., phone use), or (b) partially shared, with a limited number of channels dedicated for traffic purposes only. A partially shared system might use special data channels using communications protocols that eliminate the need for call setup and tear-down.

4. A one-way FM subcarrier broadcast can be used to send traffic information from the infrastructure to properly equipped vehicles. This is the only communication alternative considered that does not support two-way communications (e.g., does not support vehicle probe reports). This communication option was selected because it provides an efficient and relatively low-cost means of broadcasting traffic information to many users. Similar to the wide-area radio system described above, an FM subcarrier offers a means by which identical traffic information can be transmitted to all cars within receiving range.

Location of the Route Selection Function

Another distinguishing feature of the architectures is the location of the route selection function. Route selection provides recommended routes that are inputs to the route guidance function in the vehicle. Route selection involves the calculation of a "best" route of travel—selected according to an objective function—between a specified origin and destination (O-D) pair, on the basis of dynamic data, static data, or a combination of both. Operationally, the route selection algorithm would have to be periodically rerun to account for changing traffic conditions.

Two alternative locations for route selection have been considered for the strawman architectures: in-vehicle and in the infrastructure. The additional assumption has been made that infrastructure-based route selection is performed centrally at a single facility (e.g., the TMC). This latter assumption is important to the concept of coupling between route selection and traffic control, as described in the next section.

In fully distributed route selection the function occurs within individual vehicles (but without coordination with other vehicles). The driver requests a route and receives route guidance instructions from the in-vehicle system. When the vehicle is within a TMC's coverage area, the in-vehicle route selection processor is provided with dynamic traffic information from the TMC. This traffic information would consist of changes in link travel times, incidents, and other temporary restrictions (exception-type data). The in-vehicle system could also operate without reliance upon external (infrastructure) support when outside a traffic information coverage area or when dynamic traffic information is not available. In this case, traffic information used for route selection would consist of static information stored in the vehicle's map data base.

In contrast, with centralized route selection, individual vehicles transmit specific origin (or current location) and destination data to the infrastructure. This information is then used to select routes for the equipped vehicles that have provided O-D information. Each of these vehicles receives individualized route advice to support its trip. The vehicles will periodically send updates of their current location and, if necessary, receive new routing instructions for the remaining set of road links to be traveled. A vehicle might or might not utilize a map data base to assist the display of the selected route to the traveler.

Degree of Coupling Between Route Selection and Traffic Control

Coupling refers to the extent to which route selection and traffic control decisions are interdependent and coordinated. We consider three alternative levels of coupling:

- Fully coupled: TMC has real-time knowledge of equipped vehicles' O-D information and simultaneously optimizes both traffic controls and the routes of all equipped vehicles;
- Uncoupled: Traffic control optimization is carried out without real-time knowledge of vehicles' O-D and route information; and
- Partially coupled: TMC uses real-time knowledge of the routes selected by equipped vehicles, as well as their O-D information, for optimizing traffic controls.

With a fully coupled system, the TMC jointly determines the optimal settings for all traffic control devices and the optimum routes for all vehicles equipped to carry out route guidance. The capability requires complete knowledge of the origins (or current locations) and destinations at the TMC of equipped vehicles, which would be used with other traffic data obtained from traffic sensors and vehicle probes. Figure 2*a* provides a simple representation of the joint optimization process. The TMC would concurrently perform several functions:

- Prediction of future traffic loads and link times on the road network;
- Optimization of settings for traffic control devices; and
- Route selection for all equipped vehicles.

In an architecture with no coupling, the vehicles do not provide either destination or intended route information to

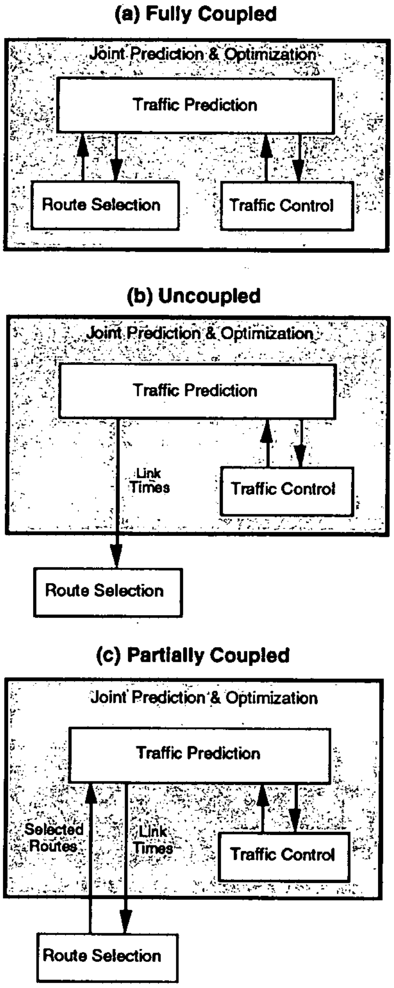


FIGURE 2 Degree of coupling between route selection and traffic control.

the TMC. The TMC only uses real-time information from traffic sensors, including vehicle probes, and historical traffic data to predict traffic loads on the network. It then optimizes the traffic control system on the basis of these (limited) traffic predictions. The link times from this process are passed to an independent route selection processor. The traffic control settings are not affected by information about selected routes; however, the traffic control optimization can affect the route selection process (through the link times). Figure 2*b* characterizes the traffic control and route selection functions for an uncoupled architecture. With no coupling, route selection could occur in either vehicles or the infrastructure.

With partial coupling, the TMC obtains planned destinations and selected routes from equipped vehicles. This information differentiates the partially coupled architecture from an uncoupled one. The TMC uses traffic sensor and vehicle probe information to estimate current link travel times. Using these inputs, as well as the selected routes and destinations, the TMC can predict traffic loads and optimize traffic control systems. Figure 2*c* provides a representation of the route selection and traffic control optimization process for a partially coupled control architecture. Partial coupling can exist with route selection in either the vehicles or the infrastructure.

Other ATMS and ATIS Functions

Although the development of the strawman architectures focused on three key issues, the architectures include several additional ATMS and ATIS functions. Except for the level of coupling of route selection and traffic control, each architecture is assumed to be able to provide the same traffic management functions. The main ones are as follows:

- Surveillance,
- Traffic data preprocessing and fusion,
- Traffic assignments and prediction,
- Dynamic traffic control (i.e., signal and sign control),
- Incident identification, and
- Traffic and safety advisories.

The ATIS services that are provided by the architectures are as follows:

- In-vehicle trip planning (excluding dynamic public transit information),
- Position location,
- Route selection,
- Route guidance for the selected route,
- Assistance requests, and
- Traveler information services.

THE STRAWMAN ARCHITECTURES

Five strawman architectures have been selected that reflect logical combinations of the three key architectural attributes described above. Table 1 gives the salient features that distinguish the architectures from one another. Each alternative provides a different framework upon which required IVHS features and services can be assembled.

The first four architectures represent highly capable ATMS/ATIS architectures. Although these architectures differ in important ways, they all have essentially the same basic capabilities; specifically, they all include some form of vehicle routing, two-way communications, advanced traffic management, and so forth. The fifth alternative represents an architecture with less functionality, primarily because it does not provide an inbound (vehicle-to-infrastructure) communication link.

A descriptive overview of the architecture alternatives is provided below. Additional details can be found in earlier work (*3*). Figure 3 highlights for each of the strawmen the location of the route selection function, the primary functions provided, and the communication medium between the vehicle and the traffic management infrastructure. In Figure 3, the components of each architecture are classified into three categories:

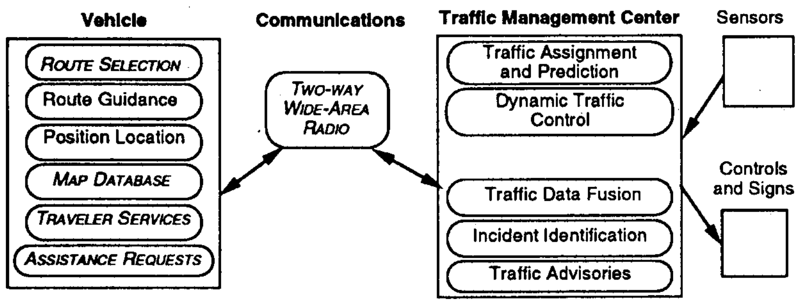
- In-vehicle,
- Traffic management infrastructure, and
- Vehicle-infrastructure communications.

Strawman Architecture 1

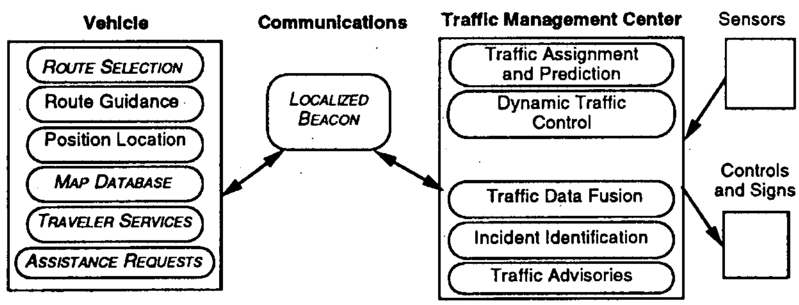
Architecture 1 has route selection and guidance capability in the vehicle with no coupling between vehicle routing and the

TABLE 1 Overview of Strawman Architectures

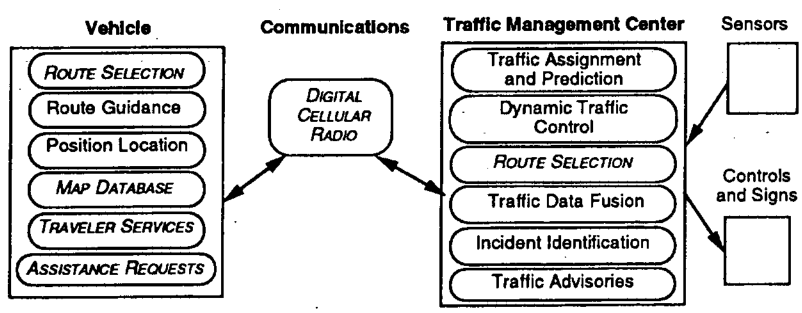
Architecture 1	Architecture 2	Architecture 3	Architecture 4	Architecture 5
Uncoupled route selection/traffic control	Partially coupled route selection/traffic control	Fully coupled route selection/traffic control	Uncoupled route selection/traffic control	Uncoupled route selection/traffic control
		Centralized route selection (with real-time data)	Centralized route selection	
In-vehicle route selection	In-vehicle route selection	In-vehicle route selection (with static data) when real-time routing is unavailable		In-vehicle route selection
In-vehicle map database	In-vehicle map database	In-vehicle map database	No in-vehicle map database	In-vehicle map database
Two-way wide-area radio communications	Two-way localized (beacon) communications	Two-way cellular radio communications (digital technology)	Two-way localized (beacon) communications	One-way broadcast (FM subcarrier) to vehicles for communications



(a) Strawman #1

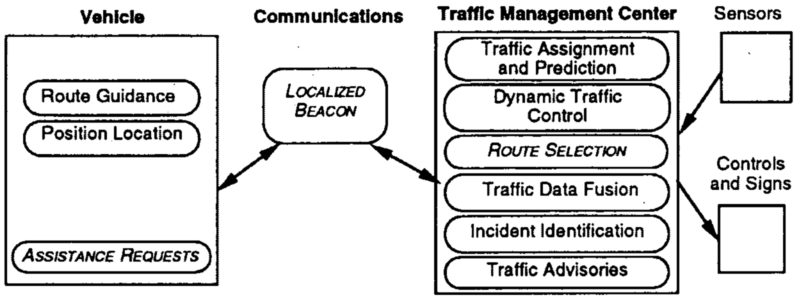


(b) Strawman #2

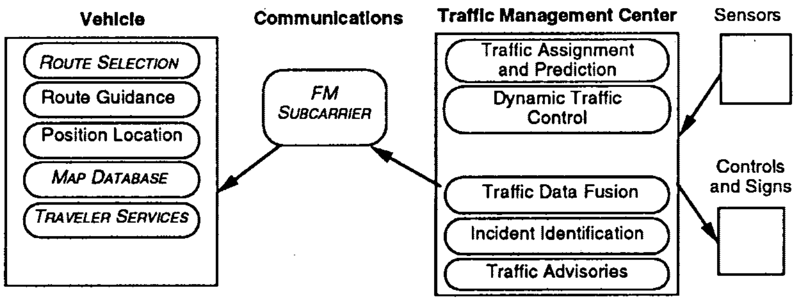


(c) Strawman #3

FIGURE 3 Simplified diagrams of strawman architectures. *(continued on next page)*



(d) Strawman #4



(e) Strawman #5

Notes: *ITEMS IN ITALICS AND ALL CAPS, AS IN THIS SENTENCE, VARY AMONG THE ARCHITECTURES.*

ROUTE SELECTION varies in its location, i.e., whether it is performed in the vehicle or in the Traffic Management Center. (Also in Architecture 3, in-vehicle *ROUTE SELECTION* uses static data only.)

The choice of the *COMMUNICATIONS MEDIA* varies across architectures.

The (in-vehicle) *MAP DATABASE*, and the *TRAVELER SERVICES* and *ASSISTANCE REQUEST* functions are included in some architectures and omitted in others.

FIGURE 3 ***(continued)***

traffic control system. The in-vehicle system performs position location, route selection, and enroute guidance functions. Position location within the vehicle is based on land-based or satellite radio trilateration, dead reckoning, and map matching. The in-vehicle navigation and routing unit contains a map data base that supports traveler services and "yellow pages" data. A variety of trip planning and route selection options are available in-vehicle to support the traveler.

The in-vehicle unit communicates with the TMC via a wide-area radio system that may cover a metropolitan area, critical segments of Interstate highway systems, or other rural areas.

Strawman Architecture 2

Architecture 2 is similar to Architecture 1, with some important modifications. Like Architecture 1, all position location and routing functions are performed in the vehicle and all high-level traffic control functions in the TMC. However, Architecture 2 has partial coupling between the traffic control and routing functions. This level of coupling is accomplished by the transmission of selected routes to the TMC, where they are used to enhance the accuracy of traffic prediction and traffic control functions. In the TMC, the selected route data must be integrated (fused) with other traffic information.

A major difference between Architectures 1 and 2 is in the type of vehicle-to-TMC communications. This architecture uses two-way localized communications (beacons) instead of a wide-area network. This beacon system must have the capability of transmitting traffic data describing link times on all major roads in the network to all vehicles that have route selection capability. This will require large messages to be transmitted in a short time by the beacons.

Strawman Architecture 3

This strawman has full coupling of the traffic management and routing systems. The simultaneous determination of the traffic control and routing parameters requires that communication between the two processes take place frequently and efficiently. This is best accomplished if both processes take place in a single physical facility, namely a TMC. Therefore, this strawman has the selection of traffic routings determined centrally, rather than in individual vehicles.

Recommended routes are communicated to the equipped vehicles in Strawman 3. This is a major difference from Strawmen 1 and 2, in which link times are sent to equipped vehicles. For this strawman, the communication link is provided by two-way digital cellular radio. The routing data for each vehicle must be coded with the vehicle's identification (ID) so that the vehicle can select out the messages addressed to it.

An in-vehicle route guidance processor utilizes the recommended routes sent from the TMC to provide assistance to the traveler about directions and turns. Vehicle location is determined by the individual vehicles, and each vehicle is assumed to be able to perform in-vehicle navigation when routing recommendations are not available from the central facility.

Strawman Architecture 4

Architecture 4 has centralized route selection with an uncoupled traffic control system. Similar to Architecture 2, localized beacons are used for communications between the infrastructure and vehicles. In this architecture, communication from the infrastructure to the vehicles is required for any route guidance to be displayed, even using static data. With this centralized approach, a map data base is not required in the vehicle, potentially lowering the cost of the in-vehicle unit. Thus, Architecture 4 is highly infrastructure dependent.

Even though both route selection and the determination of control system parameters are carried out centrally, there is no direct coupling between these two functions. These two functions do not have to be carried out at the same facility and conceivably could be handled by different organizations. For example, route selection could be performed by a private subscription service while traffic management is provided by a public agency.

Strawman Architecture 5

Architecture 5 has uncoupled route selection and traffic control with a one-way, FM subcarrier communication system to broadcast real-time traffic conditions from the infrastructure to vehicles. The broadcast is assumed to transmit in digital format. The same broadcast can be used for transmitting both routes and traffic information. The traffic management functions of this architecture are similar to those of Architecture 1, except that no vehicle probe data are collected.

The communication infrastructure is assumed to be developed primarily to provide information to vehicles with route selection and route guidance processors but also allows vehicles without these routing processors to obtain some of the transmitted information. Both equipment options can be accommodated with one overall architecture and are intended to operate together in the same area. In addition to carrying general traffic advisories and warnings, the one-way traffic channel is used to broadcast link travel times that can be used for routing purposes. The main limitations of this architecture, compared with the previous four architectures, are that it does not support assistance requests or vehicle probe traffic data.

INITIAL EVALUATION OF THE STRAWMAN ARCHITECTURES

The evaluation compared the capabilities of the five strawman architectures according to several important criteria. The evaluation was qualitative and relied on currently available information; however, it provides a structure that can be used as more quantitative information becomes available. Evaluation criteria that might be pertinent only for comparing detailed designs were generally avoided. The qualitative evaluation criteria that were used were chosen because they identify important characteristics of the systems and can provide initial insights about system benefits and costs.

The categories of evaluation criteria that were used to evaluate the strawman architectures are as follows:

- System performance (end state),
- System costs (end state),
- System risks,
- System evolution, and
- Institutional issues.

The evaluation focused on characteristics of the alternative architectures that affect performance, cost, and risk. Because the five architectures being compared are end-state architectures, the system performance and cost criteria that were used relate to the characteristics of fully developed systems. In using these criteria, it was assumed that the relevant technologies have matured to the (end) state implied by the architectures and that the market penetrations of the systems have grown to their highest (saturated) levels. Several criteria were chosen that relate to the technical, operational, and institutional risk of the architectures. These criteria address the risk of services or systems not being provided, as well as the risk of services not being efficiently utilized by either traffic managers or vehicle operators.

As part of the evaluation, several issues related to system evolution were also considered. These are connected with selecting a deployment strategy that could produce a desirable end-state IVHS architecture from feasible initial start-up conditions. Institutional concerns, such as equity and privatization, also were investigated. The major geographic focus of the evaluation was on large metropolitan areas because the strawman architectures emphasize route selection and traffic management functions.

Attempts have been made to give each of the alternative architectures a ranking for each of the qualitative criteria. A ranking is used as an ordinal comparison: an architecture has first-, second-, or third-level rankings, based on its relationship to other architectures. The detailed methodology used to obtain the rankings is described in the related report (*4*).

Because there are several evaluation criteria, each strawman architecture received several rankings. The rankings have not been combined into overall scores for the alternatives. However, a summary of the architecture rankings is provided in Table 2. Caution must be used when interpreting the results because several important criteria were not rated and because of the many assumptions made for those criteria that were rated. Also, an ordinal ranking gives no information on the magnitude of differences. This means that one is not able to estimate how closely, say, a second-ranked alternative is to a first.

Despite the incompleteness of the evaluation, a few generalizations can be derived from the rankings. Architecture 5 rated very well for each evaluation category except for system performance, operational characteristics. Its poor rating

TABLE 2 Summary of Rankings for Alternative Architectures

EVALUATION CRITERIA	ARCHITECTURE 1	2	3	4	5
SYSTEM PERFORMANCE					
Operational Characteristics					
Accuracy of Traffic Prediction Models	2nd	1st	1st	2nd	3rd
Efficiency of Traffic Control System	2nd	1st	1st	2nd	3rd
Efficiency of Route Calculations					
Accuracy of Position Location	1st	1st	1st	2nd	1st
Effectiveness of Information Delivery Methods					
Adequacy of Communication System Capacity					
System Reliability and Maintainability					
Capability During Communication Failure	1st	1st	1st	2nd	1st
Capability During TMC Failure	1st	1st	1st	1st	1st
Capability to Update Maps					
Performance in Poor Weather	1st	2nd	2nd	2nd	1st
Performance in Diverse Terrain/Structures	1st	1st	2nd	1st	1st
SYSTEM COSTS					
Vehicle Capital	3rd	3rd	3rd	1st	2nd
Vehicle Operating	3rd	3rd	3rd	1st	2nd
Communication System Capital	2nd	3rd	2nd	3rd	1st
Communication System Operating					
Traffic Management Capital	1st	2nd	3rd	2nd	1st
Traffic Management Operating	1st	2nd	3rd	2nd	1st
SYSTEM RISKS					
Equipment or Service Not Being Provided					
Communication System Risks					
Other System Risks	1st	2nd	3rd	2nd	1st
Service Not Implemented by Traffic Managers					
Complexity of TMC Risk	1st	2nd	3rd	2nd	1st
Failing to Meet Jurisdictional Interest Risk	2nd	2nd	1st	1st	2nd
Service Not Purchased or Used by Vehicle Operators	2nd	2nd	3rd	2nd	1st
Technology Places Limits on Size of Market					
Outbound Communications	1st	1st	2nd	1st	1st
Inbound Communications	2nd	1st	1st	1st	N/A

Note: (1st = best ranking for each criterion, e.g., best performance, lowest cost, lowest risk)

in the operational category can be attributed to its lack of inbound communication capability. Architectures 2 and 3 rate similar to one another, largely because they are the only architectures with some level of coupling. Architectures 1 and 4 also appear to rate similarly for the evaluation categories other than system costs.

CONCLUSIONS

Several conclusions were reached that, although based on qualitative analyses, generally appear to be well founded. The conclusions relate to both the key attributes of the architectures—route selection, traffic control, and communication—as well as to the complete architectures. Some ways to mix the architectural attributes to produce promising new alternatives are suggested, and several unpromising combinations of architecture attributes are identified. Some important observations about system evolution are also discussed.

Route Selection and Traffic Control

The location of route selection cannot be decided in isolation from other architectural decisions. The location may not directly affect the efficiency of selected routes but certainly has an impact on it—through the level of coupling between route selection and traffic control. If full coupling can be accomplished, collocating route selection with the traffic prediction and control functions and using system optimization may be the preferred approach.

Coupling of route selection and dynamic traffic control in an integrated process has potentially high payoff, but there is currently great uncertainty about coupling's technical and institutional feasibility. The technical uncertainty is related to the accuracy of traffic prediction models and the ability to provide needed processing requirements. Coupling places stringent requirements on all parts of the system architecture for very fast response times, including the communication system and the traffic sensors.

Institutional uncertainty is related to acceptance by traffic managers and vehicle operators. Traffic managers may be concerned about the overall complexity of the computer algorithms or the lack of manual control that they would have. Travelers may choose to ignore system-optimized routes that do not consider user preferences. This potential concern of users can be important even though the alternative—placing the route selection processor in the vehicle—would likely cost the user more. (This last observation might not hold if central routing information providers charge hefty fees to users of their routes.) Centralized systems can optimize routes for each user, rather than for the overall system, but then the potential benefits of full coupling would be lost.

Partial coupling may provide an approach that is acceptable to vehicle operators (because it has user-preferred routes) while also being acceptable to traffic managers. But partial coupling requires that traffic predictions be able to factor in user-selected routes quickly enough to change the estimated link times before a large number of vehicles make the same route diversions. Otherwise, the result could be instability of the traffic control system (thrashing).

Dynamic traffic control is an important component of all the strawman architectures. To be successful, it has to work well for most ATIS and ATMS services. The details of this function have not been examined in this evaluation, but one major aspect—traffic assignment and prediction—has been considered. The ability to make, and continuously update, predictions of traffic flows and link times for several minutes into the future using real-time data is a major requirement for providing dynamic traffic control. This requirement, in turn, depends on acceptable computer hardware and software, as well as a sufficient number of traffic sensors sending current data and a sensor-to-infrastructure communication system that has the speed and capacity to send frequent updates.

The more timely and accurate the traffic information that can be obtained, the better will be the traffic predictions. For this information to be truly valuable, however, the prediction function must be able to estimate traveler responses to the routing advice. One way for this to occur is to have a traveler behavior model embedded in the prediction model.

In all the strawman architectures, the traffic management processing function has been discussed as if it were performed in a single centralized facility. This was done for ease of comparison with in-vehicle processing. In fact, the processing connected with traffic management could be distributed among several centers; for example, one metropolitan center and several subregional ones. In this case, there would be additional architectural issues about how the optimization is coordinated, as well as by what means and how frequently information is transferred between centers. Of course, coupling with route selection would make more difficult any distributed processing of the traffic management requirements. Another variation of decentralized traffic management would be to carry out some of the tasks on the roadside in localized facilities. Again, coordination of the traffic control optimization would be an issue. However, it is likely that there would be some form of hierarchical control system with local controllers that function independently, within constraints set by a network-wide optimization plan.

Communications

No definitive conclusions were reached about the various communication systems in this initial evaluation. Each has different advantages and disadvantages that cannot be compared in a qualitative analysis. However, a quantitative analysis of communication load requirements has recently been conducted at MITRE (*6*) that provides more insight into this subject.

A major question about each architecture is the cost allocation of the communication services among the various owners of the IVHS components. Because service providers can charge user fees, the answer to this question will be resolved only when a particular implementation of the service is examined.

End-State Architectures

With regard to the five end-state architectures, but with all communication systems combined with the other key attributes, several conclusions can be reached. First, the requirement of passing link data to vehicles can best be met with a broadcast channel. This is a fairly weak conclusion, because cellular radio may use dedicated broadcast channels, and beacons may have a high enough data rate, when used with a structured flow of traffic, to also be able to transmit link data without practical problems.

It has also been concluded that passing selected routes from a traffic management facility to vehicles should not be performed by a wide-area broadcast system, because there is not a good match between the large data transmission requirement and channel capacity. The broadcast channel would have to send all routes to all vehicles (although each vehicle could pick off only the data that it needed).

An inbound broadcast channel, such as that provided by wide-area radio, does not appear to be a good match for sending selected route data from vehicles to a TMC, such as is required with partial coupling in Architecture 2. This is because of communication system access limitations, which, even with a stringent access protocol, could prevent some routes from being sent. These observations imply that wide-area radio, by itself, should not be considered for Architectures 2, 3, or 4. (In fact, it was considered only for Strawman 1.) It also appears that an inbound broadcast channel is not the best mode for sending probe data.

Another conclusion is that an architecture that supports infrastructure-based route selection without an in-car electronic map should include a beacon type of communication system to serve as location fixes. Although it would be possible to design a mapless system without beacons, it does not appear to be a promising approach.

It may have been observed that Architecture 4 appears to have route selection, as well as traffic management, occurring in a central traffic management facility, while not having these two functions coupled together. The original idea behind the strawman was that there would be a desire to carry out route selection centrally by traffic managers most likely on the basis of system optimization. But there would be a constraint, either institutionally or technically, in coupling with traffic control. This constraint might occur because the two functions are carried out in two different facilities, or by two different organizations, or even because the traffic management is carried out in a distributed manner among several levels of TMCs. Although it has not been possible to consider any of these variations, at least some of them need more investigation.

The utility of each of the architectures for the APTS, CVO, and AVCS service areas was examined briefly in the related report (*4*). One conclusion is that it is generally possible to have separable architectures for these additional services—

even for AVCS services in a vehicle that also performs the route guidance function. In addition, the communication requirements for ATIS services that are performed outside the vehicle appear to differ from those that are performed within the vehicle. When a full range of IVHS services must be provided, it may turn out that multiple communication systems must be utilized. One exception is that, within some frequency ranges, a broadcast to vehicles could also be received inside buildings. FM subcarrier, for example, fulfills this requirement.

System Evolution

The potential for evolutionary development of an architecture is one of the most important issues to consider in an evaluation, especially for a system such as IVHS. Because of the technical and institutional uncertainties of the IVHS program, an end-state architecture should be defined so that it can "gracefully" evolve from a simpler, less costly, and less risky one. An architecture that provides a range of services and functions could be developed to make some functions available in the near term, whereas others would be added later. This evolutionary approach would allow improvements in performance over time, as the number of users increases. A start-up architecture should be conceptualized to provide significant initial benefits, while being capable of evolving without the need to discard major system components.

One attractive start-up architecture is Strawman 5, which has only a one-way communication channel; however, it includes a capability to broadcast traffic link data to vehicles that have in-vehicle route selection equipment. This strawman has little risk and relatively small cost, but is does not provide the same level of functionality as the other architectures. However, this start-up architecture does not have to be a dead-end architecture, because an additional vehicle-to-infrastructure communication link could be added to provide for the transfer of either probe data, as in Architecture 1, or probe and route data, as in Architecture 2. The system would have to be designed to allow for easy in-vehicle equipment upgrades for users that wish to add the additional communication functionality. An in-vehicle route selection approach would be maintained as the system evolved.

Another evolutionary possibility focuses on having centralized route guidance along with system optimization in the end state, as in Architecture 3. The start-up architecture could be a version of Architecture 4, which has no coupling, and could be based on a cellular or a beacon communication system. It is not known at this time which communication system would require a higher start-up cost.

It also may be possible to start with a capability built around in-vehicle route selection and evolve to centralized route selection with system optimization. The details of this evolution are not completely clear. The physical components of the communication capability, if they were based on beacons or cellular, might not have to evolve appreciably. However, the information exchange protocols would have to be modified (if the end-state capability were not already provided for in the beginning). In the end state, vehicles would continue to utilize map data bases but would use in-vehicle route selection only with static data, for example, when centrally determined routes are not available. The TMC would have to make major improvements to its hardware and software to provide the new capabilities.

Institutional Issues

Some preliminary observations can be made about two important institutional issues that should be important evaluation concerns: equity and privatization. In considering the costs of both alternative communication equipment and of the data that could be made available to vehicles, it has been determined that there are likely to be fees charged by either the communication operators or the information providers. These fees may be set only to transfer the cost from a service provider to a service user. But there is also the possibility that fees will be set to accomplish other ends, such as profitability, equity among various travelers, pricing low to buy into a market, and so on. In other words, the fee structures can easily dominate considerations of equity of benefits and costs of IVHS services.

The type and size of fees will depend on which IVHS services are operated publicly and which ones are operated privately. Hence, the issues of equity and privatization should not be dealt with independently. However, they are not architecture issues per se; they are implementation issues. There can be a range of public and private roles for each of the strawman architectures that has been considered so far. Hence, it appears that the level of privatization, by itself, cannot be used to discriminate among architectures.

ACKNOWLEDGMENTS

The work described in this paper was performed under an FHWA contract. The authors would like to thank Michael McGurrin, John Cerva, Allen Mertig, and Vijay Patel of The MITRE Corporation for their technical advice and contributions to this work and the members of the IVHS AMERICA System Architecture Committee for the very valuable comments they provided on the work as it progressed.

REFERENCES

1. *Strategic Plan for Intelligent Vehicle-Highway Systems in the United States*. Report IVHS-AMER-92-3. IVHS AMERICA, Washington, D.C. 1992.
2. A. Yablonski. *Working Paper on IVHS User Services and Functions*. MP-92W043. The MITRE Corporation, McLean, Va., 1992.
3. M. D. Cheslow, S. G. Hatcher, M. F. McGurrin, and A. A. Mertig. *Alternative IVHS Architectures*. MTR-92-W-048. The MITRE Corporation, McLean, Va., 1992.
4. M. D. Cheslow, S. G. Hatcher, and V. M. Patel. *An Initial Evaluation of Alternative Intelligent Vehicle Highway Systems Architectures*. MTR-92-W-063. The MITRE Corporation, McLean, Va., 1992.
5. K. Chen and S. Underwood. Research on Anticipatory Route Guidance. *Proc., IEEE Conference on Vehicle Navigation and Information Systems (VNIS)*, 1991, pp. 427–439.
6. M. D. Cheslow and S. G. Hatcher. Estimation of Communication Load Requirements for Five ATIS/ATMS Architectures. *Proc., 3rd Annual Meeting of IVHS AMERICA*, Washington, D.C. (in press).

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Intelligent Vehicle Highway System Safety: Specification and Hazard Analysis of a System with Vehicle-Borne Intelligence

A. HITCHCOCK

Hsu has described the normal operation of a system of automated freeways that minimizes the degree to which the infrastructure is involved in maneuvers. No account is given of procedures on entry and exit or of possible faults. The Partners for Advanced Transit and Highways (PATH) safety program demanded a second example of the process of full specification and fault tree analysis to determine if this process was generally applicable. Accordingly, Hsu's specification has been completed, retaining minimal infrastructure-based involvement in the maneuvers. In this design the safety criterion was adopted that no hazard shall arise unless there are two independent faults. The hazards were those used earlier. Because a hazard is the precursor of a catastrophe, no multivehicle high-relative-speed collisions should occur without three independent faults. A fault tree analysis was carried out in which there were never more than four branches in any line of the tree. It is concluded that it is possible to produce a design that meets these safety criteria and that behaves during normal operation in the same way as the system defined by Hsu. Further, the method of complete specification and fault tree analysis suffices to produce a system of demonstrated safety with a practical allocation of resources and time. A comparison is made of the safety of this vehicle intelligence system and the earlier infrastructure intelligence one. On the basis of experience rather than analysis, it is concluded that it is possible to design a safe automated freeway.

The process of complete specification and hazard analysis has been defined previously in relation to automated freeways by Hitchcock (*1*). The procedure already has been demonstrated in one example (*2–4*). That was a case in which the system intelligence was mainly located in the infrastructure. The fault tree analysis was successful in finding errors in the design. There is interest also in evaluating cases in which the intelligence is mainly vehicle based. It is possible that the basic method would be less successful or would need modification in such a case. Therefore a second example was undertaken of complete specification and fault tree analysis, in which the system considered is based on vehicle-mounted intelligence (*2,5,6*).

First it was necessary to define the hazards. A hazard is defined as a situation in which a further fault can lead to a catastrophe. A catastrophe is defined, as is usual in system safety, as a multicasualty event; in practice this means a high-speed collision involving platoons or vehicles. The hazards restated below are the same as those used in the earlier work (*1,2*). As in the previous work, the statement of hazards applies to platooned systems. However, in principle they apply to nonplatooned systems also, although small changes to the formal statements are required. Also, for the reasons given by Hitchcock (*1*), there are barriers, called dividers, between the lanes that prevent vehicles from straying from one lane to another. The hazards do not expressly refer to this, although they can again be construed to apply to the problems that arise if the dividers are absent. For example, if a vehicle strays into another lane, it becomes the "one ahead" in Hazard 1 below. The hazards are as follows:

1. A platoon (or single controlled vehicle) is separated from one ahead of it, or from a massive stationary object in its path, by less than platoon spacing (see below).
2. A vehicle not under system control is an unmeasured and unknown distance in front of a platoon or free agent (or single controlled vehicle).
3. A vehicle is released to manual control before the driver has given a positive indication of readiness.
4. A vehicle is released to manual control at less than manual spacing from the vehicle ahead of it or at such a relative speed that a spacing less than manual spacing will be realized within (say) 2 sec.

A platoon is a succession of any number of vehicles that are closely spaced in the same lane and under coordinated control. In the hazards above, platoon spacing is defined as the safe spacing between platoons according to the criterion of Shladover (*7*). Manual spacing is the spacing at which drivers feel comfortable and that they use in normal (i.e., not automated) driving. In the system proposed, these quantities, which depend on the condition of the road, are set by the system controllers.

To say that a system meets, or does not meet, a safety criterion is a statement of fact. This paper discusses means by which the validity of such a statement can be demonstrated. To say that a system is safe, on the other hand, is a statement that will mean different things to different people.

California PATH Program, Richmond Field Station, Institute of Transportation Studies, University of California, Berkeley, Berkeley, Calif.

SYSTEM SPECIFICATION

When the work reported here started, a system with the requirements for this work had been partially specified by Hsu et al. (8). This system was used as the basis for a complete specification. A full description is given by Hitchcock (5). Although the full specification is much too lengthy to be reproduced here, in this paper the specification and the design method are described and examples are given.

It is a principle of the design method adopted here that procedures such as checking for faults are integral to the design. To design a system for normal operation and then "bolt on" safety features has been shown by experience in other fields to lead to difficulties. The designer therefore considered what checks and messages were necessary to confirm compliance with expectations at each stage, both with and without a detected fault. In the description that follows, system behavior in normal and fault conditions is considered in parallel. The architecture encompasses both functions.

Each length of lane in the system (called a block, defined below) may operate in one of several modes. In some of these, vehicles are at rest while emergency vehicles under human direction deal with accidents. One is a restart mode. One is the normal operating mode, and the others are degraded ones, in which speeds, for example, may be reduced to facilitate lane changing if a lane is closed ahead (SA mode = slow ahead). Another example is no-entry mode (NE), used when a vehicle in the lane has a communication fault and therefore cannot cooperate if another vehicle is about to enter its lane.

As will be seen, the control system requires that messages be exchanged between vehicles. It is convenient to give such messages names that indicate their function. Such names contain an underline character (_).

Architecture

The system architecture, as described by Varaiya and Shladover (9), is a hierarchy of six levels (see Figure 1). The three lowest form the safety-critical subsystem. As safety demands, the construction is modular, and messages or data pass between modules only at fully specified interfaces. The physical level defines the physical processes (such as tire-road friction) as a result of which control actions are translated into vehicle movement. The regulatory level controls the movement of vehicles: lateral and longitudinal control (i.e., the subsystems in each vehicle that maintain it in its lane and keep it properly spaced from its predecessor) are regulatory-level functions. The selection, from among all messages received, of messages relevant to this vehicle, is another regulatory-level function. The platoon level organizes the maneuvers whose combination enables a vehicle to proceed from its entry to its desired exit. The link level advises on the desired route to achieve the objective. Because the route of each vehicle is affected by the routes of others, link-level functions are roadside based. But link advises only; before a maneuver is initiated, platoon-level procedures ensure that it can be done without hazard. The platoon, regulatory, and physical layers thus comprise the safety-critical subsystem. Functions of the higher levels will not be discussed here.

Roadway and Vehicles

The automated freeway has several lanes, divided into blocks, perhaps 10 km (6 mi) long. They are separated by dividers that contain gaps, called gates, through which lane changes are made. On the sending side of each gate (the side from which vehicles leave) there is a "turning point"—an electromagnet or other signalling device that, when activated, gives a signal to the lateral control system of a vehicle to begin a turn, if that has been commanded. If the turning point is inactive, the vehicle will not commence a lane change. The gate advises vehicles changing lanes of the position in the block of the gate, appropriate speed limits, and the system mode of this lane. The presence of a vehicle on the receiving side will inhibit turning-point activation. Rerouting instructions may also be passed here. In fault conditions, some lanes may operate in degraded modes in which speed is reduced and lane changing may be restricted. At the commencement of each block is a marker that is used by vehicles crossing it to zero a vehicle-mounted odometer. The discrete lateral control references enable distance along the lane to be measured to ±1 m or better. Every vehicle thus contains a register containing its current absolute position to within 1 m.

Each vehicle is equipped with a longitudinal control system, a lateral control system, a self-monitor, a communication system, and an overall controller:

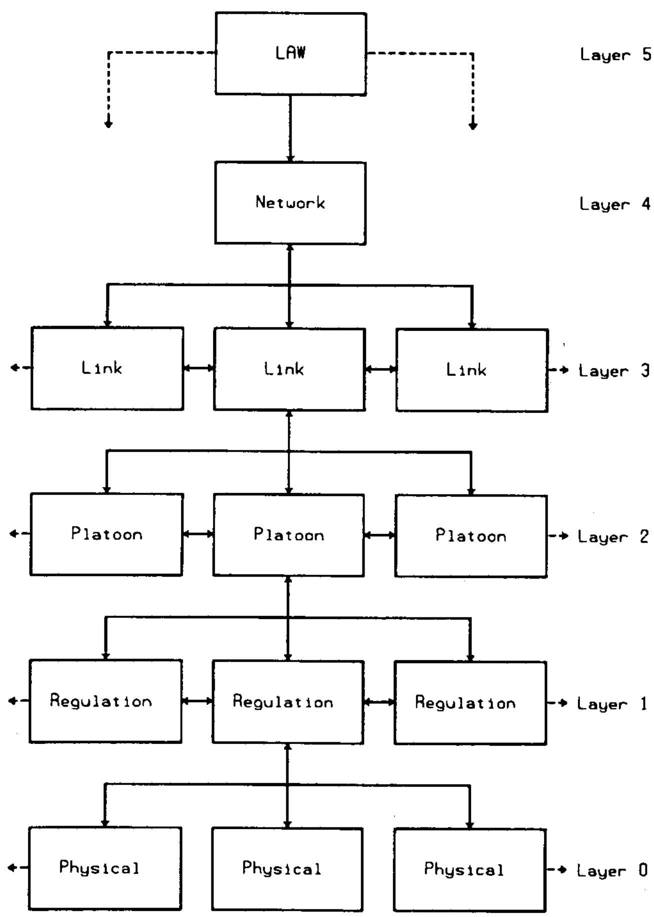


FIGURE 1 IVHS control architecture [after Varaiya and Shladover (9)].

1. The longitudinal control system has a forward sensor that can determine the distance and relative velocity of the vehicle ahead in the same lane within a maximum range. This range exceeds the maximum interplatoon spacing. If the vehicle is in platoon, the longitudinal control system will maintain the vehicle at a fixed distance (about 1 m) from the vehicle ahead. If the vehicle is a platoon leader or a free agent, the vehicle will proceed at a constant speed [between 90 and 110 km/hr (55 and 70 mph) in normal conditions], except that it will not approach closer than platoon spacing from the vehicle ahead.

The odometer provides a means for a vehicle to identify its position to others. In forming platoons, for example, the initial action is for a platoon leader to send the message request_merge (see below) to the vehicle ahead in the same lane. But the follower has no identifying name for the one ahead, and the message may be received by several vehicles. By including the odometer reading, lane number, and block number, the intended recipient can be identified.

2. The lateral control system will keep a vehicle on track in an automated lane. If a "turn" instruction has been given, it will execute a lane change at the turning point.

3. The self-monitor will detect faults in the other subsystems, including the mechanical parts of the vehicle and conditions such as shortage of fuel. If a fault is detected, the control system's objective changes. Instead of trying to follow a route designed to reach a stated destination, the vehicle will leave the automated lanes as soon as possible.

4. The communication system will transmit and receive messages from other vehicles ahead, behind, and to either side, and also messages broadcast to or from the system. These have to be eight separate independent functions: thus, for example, failure of rearward reception does not imply loss of rear transmission or of system reception. If any two of the eight functions are combined in one device, it is relatively simple to show that a hazard will arise following the single fault of its failure. This violates the safety criterion.

5. The overall controller, called supervisor, determines when it is time to join or quit a platoon, change lanes, and so on. One of its features is a busy flag, which is set while a vehicle engages in a maneuver. If the busy flag is set, the supervisor will usually deny a request from another vehicle to engage in another maneuver.

Maneuvers

Hsu et al. (*8*) describe three basic maneuvers that enable the system to perform its functions when there are no faults. In the merge maneuver, the leader of a following platoon contacts the one ahead, and, if their sizes are appropriate and the busy flag is reset, the two platoons will merge to one. In split, the reverse happens. A vehicle requests that the platoon split into two. By means of at most two such split maneuvers a vehicle in a platoon can become a free agent. Only a free agent may initiate the change-lane maneuver.

Figure 2 is based on the report of Hsu et al. (*8*) and shows the sequence of events in the merge maneuver in the present work. The appearance of an underline character (_) indicated a message name. As the figure shows, the following platoon leader, B, initiates the merge by sending a message, request_merge, to its predecessor, A. (In fact, the message is received by the rearmost member, C, of A's platoon, and passed forward by the within-platoon communications.) If A is busy, or if the resulting platoon would exceed the maximum size, A rejects the invitation by sending nack_request_merge. Otherwise A sends (via C) ack_request_merge. As it passes the message on, C sets the busy flag. B then accelerates and joins on to C. When it does so, it sends confirm_merge, again through C. Platoon leader A then changes its regularly transmitted control data message to take account of the increased platoon size. This is received successively by C and B, and as each receives the new control data, it resets its busy flag, so that it is ready to start a new maneuver.

Hsu et al. (*8*) do not distinguish the role of C, and it sets no busy flag. The change was made to deal with fault con-

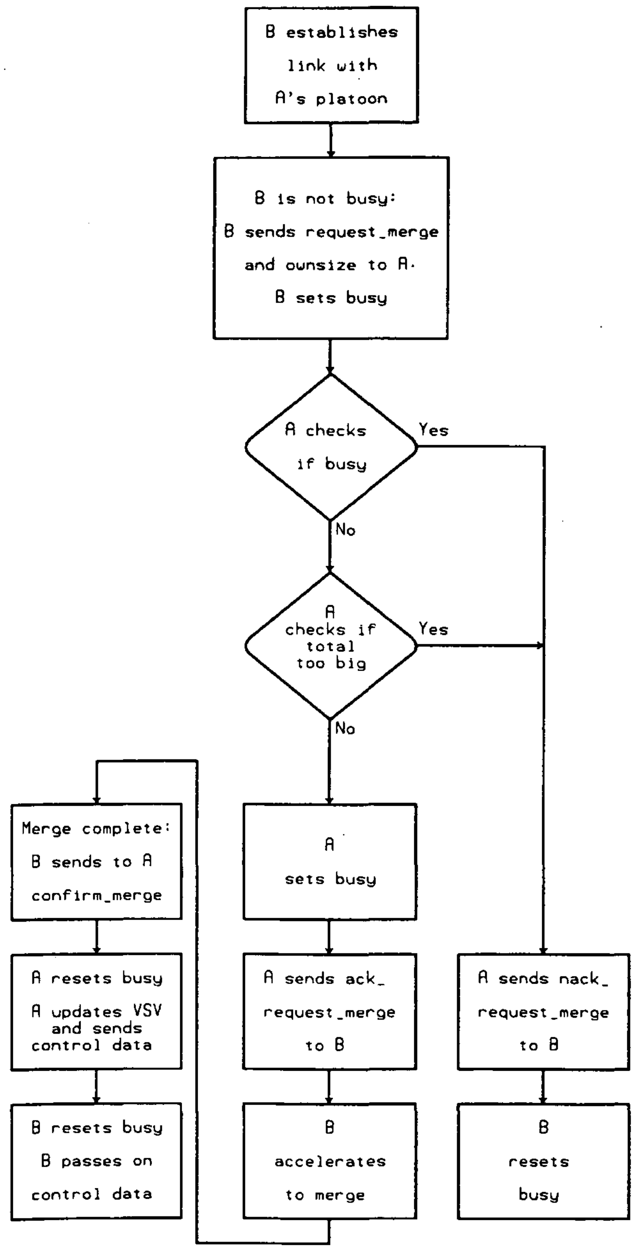


FIGURE 2 Merge [after Hsu et al. (*8*)].

ditions: C's busy flag is significant only in fault conditions. Similar changes have been made to the other maneuver protocols described by Hsu et al. (*8*). These will not be detailed here. They identify the last vehicle as part of a platoon that is engaged in a maneuver. In the case of the change-lane maneuver, arrangements are made to break off the maneuver before completion, if a forced-split maneuver (see below) is called.

Faulty Vehicles

A treatment of fault conditions was devised for this work. When the internal monitor detects a fault, or the vehicle is advised of one by the system, a flag is set in supervisor. Different flags are set for different faults. Supervisor will thereafter arrange that the vehicle (if it will still move) is driven out of the system as quickly as possible. This is necessary, as may be seen from the following argument. If a pair of unrelated faults occurs on one vehicle, a hazard may arise. Also, if a vehicle with a fault starts to interact with another, and no special action is taken, a hazard may ensue. In this system, therefore, if two faulty vehicles start to interact (because neither has yet exited), all vehicles in the relevant lanes are automatically halted. This is clearly undesirable, although not unsafe, and to avoid it vehicles with one fault should exit as soon as possible.

To extract a faulty vehicle from the system, first it must be extracted from its platoon. The extraction must be done immediately because a faulty vehicle in platoon can destroy the control arrangements for the whole platoon. If the platoon is engaged in another maneuver, the faulty vehicle will cause the maneuver to proceed to an unsatisfactory conclusion or prevent it from concluding at all. The forced-split maneuver is therefore called by a faulty vehicle in a platoon. In relevant cases it will also be called by the vehicle it has ceased to communicate with (see "probes" below). It is necessary that this vehicle call the maneuver: when communication has failed, each half must act independently. Forced split has absolute priority. If there is another maneuver in progress it is either broken off or interrupted. The busy flag in the last vehicle of a platoon ensures that a maneuver in progress is not forgotten and remains accessible. Forced split also differs from the usual split maneuver in its termination. The message confirm_for_split is sent, but no reliance is placed on its receipt. After a maximum of two forced splits a faulty vehicle becomes a free agent. Thereafter it will behave as though it were busy if a message inviting participation in a maneuver is received. Depending on the nature of the fault, the infrastructure-based part of the system will take other precautions by putting selected lanes in some blocks into degraded modes; speeds may be reduced, the lane may be cleared, or entry may be forbidden. Full details of this part of the design are not given here; the essence is that lane changing by the faulty vehicle is made both hazard free and easier. Once it is a free agent, a faulty vehicle must change lanes successively until it exits.

To do this a fifth maneuver protocol, emergency change, is required. The standard change-lane protocol may place the changing vehicle close to other vehicles or involve it in merge maneuvers with other platoons, which may be unsatisfactory and is likely to delay further changes. In the emergency-change protocol the faulty vehicle is shepherded between other platoons, if there are any in the neighborhood. Thus, even if the faulty vehicle's speed is reduced, a faster vehicle in another lane cannot interfere with the lane change.

The working of the emergency-change maneuver is described in Figure 3. As will be seen, the faulty vehicle engages the cooperation of four platoon leaders in other lanes who prevent any other vehicle from entering the change-lane area that the faulty vehicle will use. In this case, there is no need to prepare for the possibility that another emergency change or forced split will arise while the emergency-change maneuver is in progress. If another emergency-change or a forced-split is called, there is an immediate call to the system, followed by shutting down of motion in the affected lanes. Thereafter, operations will be under the control of a human on the spot (presumably a member of the highway patrol).

Probes

For most functions, on-vehicle testing will confirm or refute the presence of a fault. From time to time a situation will arise in which it is clear that there is a fault, but it is not clear in which vehicle it lies. The roadside parts of the system keep records of both faulty vehicles in their areas and "suspect" ones. (A vehicle is suspect if it is one of several that an event has shown may be faulty.) If a suspect vehicle is involved in similar incidents twice, it is declared faulty, and the other suspect vehicles are acquitted. However, there are two circumstances in which "fault" action is required, although the self-monitor cannot detect a fault. To cover these, probes are employed.

The platoon leader's probe tests the forward sensor. A platoon leader may "think" it sees a vehicle ahead either because there is indeed one present or because its forward sensor is faulty. Periodically therefore, if it is not otherwise busy, a platoon leader will send a message, indicating that it can see a vehicle ahead of it at a particular odometer reading or that it sees nothing. If it receives an unexpected reply, or no reply when one is expected, it will try again. If again there is a logical clash, the platoon leader will usually declare itself faulty. However, it will not do so if a check with the system reveals that there is a vehicle nearby that may have lost one of its communication subsystems.

The in-platoon probe deals with the situation in which a vehicle in a platoon develops a fault in communication with the vehicle behind or ahead of it. The faulty vehicle has its fault diagnosed by its self-monitor. However, the other vehicle in the noncommunicating pair must also start a forced split and must have a means of knowing that it must do so. The in-platoon probe works by means of acknowledging the control data message passed down through the platoon. Uniquely this ack_control message is not passed on to the platoon leader. The logic is shown in Figure 4.

Entry and Exit

Entry and exit are change-lane maneuvers or (for the exit of a faulty vehicle) an emergency-change maneuver. It is envisaged that there are separate on- and off-ramps. The on-ramp

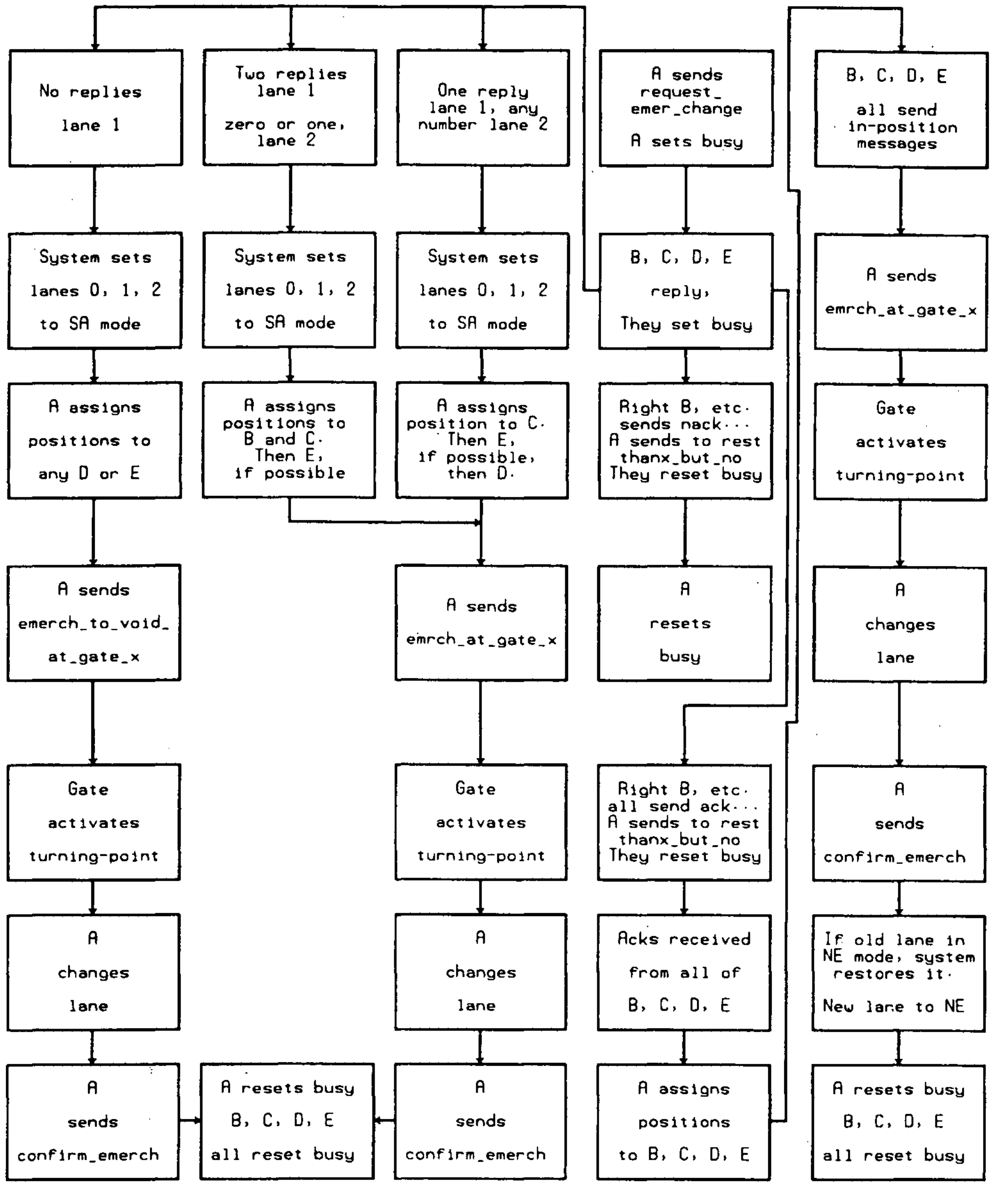


FIGURE 3 Emergency change lane. Change is to Lane 1: Platoons B and C are in Lane 1, B leading; Lane 2 is adjacent to Lane 1: Platoons D and E are in Lane 2, D leading. SA = slow ahead, NE = no entry.

is equipped with a separate exit, so that a vehicle that is refused entry can return to the manual lanes. The off-ramp is equipped with a "dormitory"—an area where a vehicle whose driver does not resume manual control on exit can be parked. On entry, a vehicle declares itself to be carrying control equipment that its self-monitor declares to be in working order. The driver must position the vehicle, relative to the traffic on the freeway, so that the change-lane maneuver is possible within the limited distance to the gate. On exit, a series of messages both before and after the change are sent to the driver and the vehicle reminding the driver of the need to resume manual control. A positive reply is needed before automatic control is released. This means, if a reply is delayed, that the vehicle must be brought to rest in the dormitory before it leaves the physical limits of the system.

DESIGN CHOICES

In the earlier work of Hitchcock (*2*) it was observed that, once the basic concepts were fixed, there seemed to be very few choices about the way the system design was put together. In the current study, there did not seem to be any design choices of importance at all, except the selection of four rather than one or two "shepherding" platoons in the emergency-change maneuver. However, it is less possible to be certain about this. The work of Hsu et al. (*8*) was taken as a basis, and there may be more choices within the area covered by that work. It is also not certain that there are no alternatives to the rule that a faulty vehicle should leave as soon as possible. Even if it is certain, it may be that there are other maneuvers besides forced-split and emergency-change, as de-

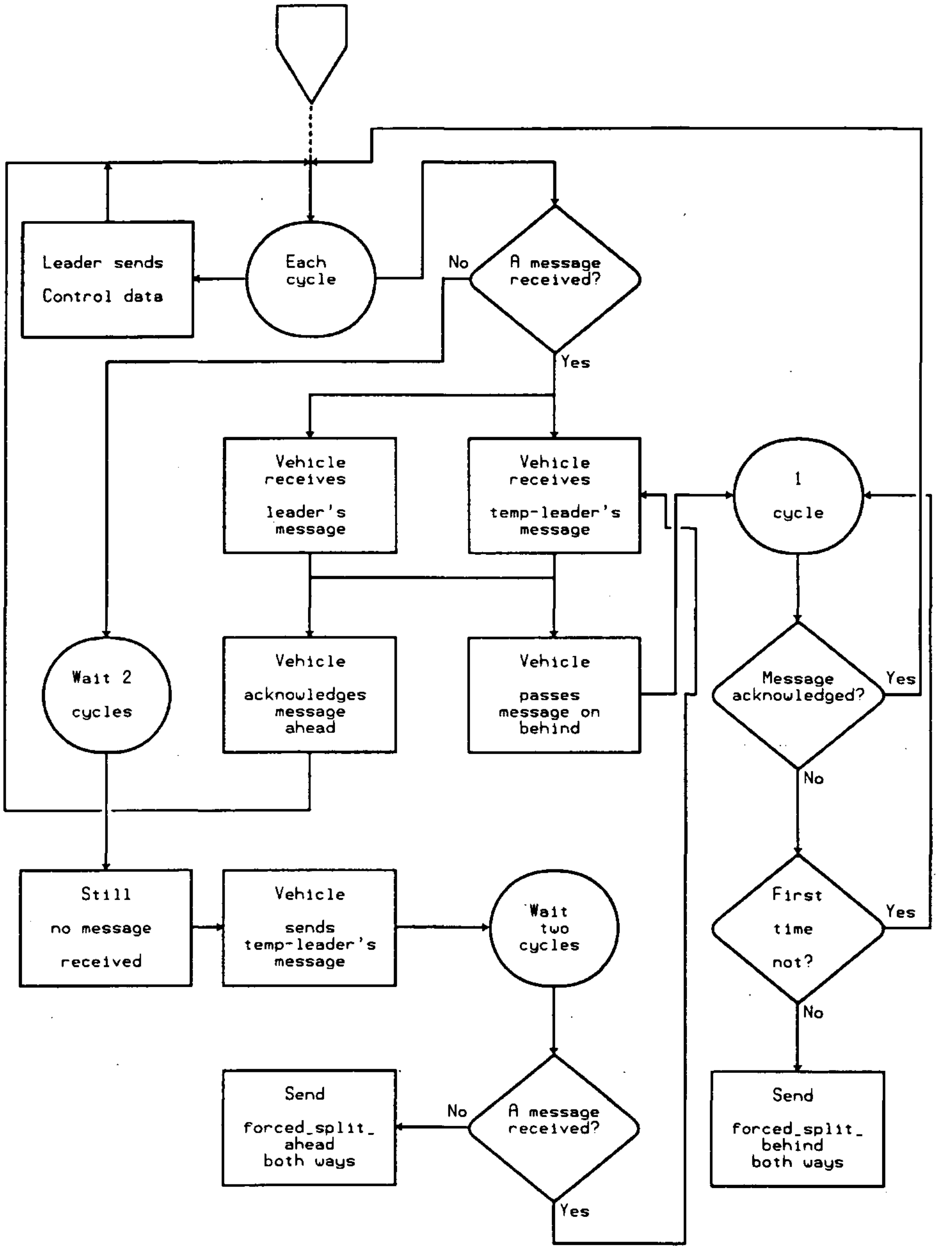


FIGURE 4 In-platoon probe. Probe operates in a cycle; there is no starting point. The figure is easier to understand by starting at the point indicated.

scribed here, that could be used to achieve this. The author did not conceive of alternatives, but this does not mean that there are none. It does, however, once again appear that the safety criteria suffice to restrict the acceptable designs to a very small number.

FAULT TREE ANALYSIS

In a fault tree analysis, each hazard is considered in turn. One asks, "How could this arise?" The answer will take on a form such as "If A happens, or B happens, or C happens . . ." One then asks "How could A arise?" "If AA happens or AB happens . . ." The process of identifying precursors continues. Mathematically, "A happens," "B happens," . . . are logical propositions, and "and," "or," and "not" are Boolean operators. Sooner or later one arrives at the point where the proposition is one of the following:

1. "This" (A, B, C, etc.) can happen as a result of a single fault in a vehicle or other system component. In this case a design error has been found.
2. "This" implies that two simultaneous faults have occurred.
3. "This" implies that there has been a computer error (the computers are assumed to be so redundant that this implies two simultaneous faults).

4. "This" is a proposition and is not possible (e.g., involves reversal of gravity).
5. "This" is a proposition that implies that there has been systematic failure to maintain the infrastructure.

In each of the last four cases there is no breach of the safety criterion on this branch of the tree.

A fault tree clearly involves subjective elements. It is always possible that the investigator will fail to realize one of the ways in which a situation could arise. This becomes more likely when the investigator is the designer.

Nevertheless, in both specification and analysis, the process has been carried out with formal rigor. Each module in the design (there are about 230) has been specified in a standard form and is also included in some 15 pages of flowchart drawings. This form shares many features with the forms used for module specifications in formal-method computer languages. The specification language used here, however, is not based on formal axioms. The complete formal specification has been stated and discussed previously (*5*). In the fault tree analysis, similar rigor has been employed: there are some 60 elements in the tree, and the arguments in each have been recorded precisely (*6*). Both reports are long and complicated, and no attempt is made to summarize them here.

DESIGN FAULTS AND RESULTS OF THE FAULT TREE ANALYSIS

In no case did a branch of the fault tree contain more than three elements, which means that the fault tree analysis is practical. The alternative of arguing forward from possible combinations of faults would have involved millions of times more cases, if not billions or trillions.

The analysis detected faults of two kinds. There were some potential violations of the safety criteria, strictly construed, which were known about before the fault tree analysis was started. That the analysis detected all these is reassuring but should not convince anyone that it will detect all faults. Some faults that arose from genuine mistakes in the design process, and therefore were not foreseen, were also detected. This should increase confidence that the method will detect all errors but does not prove it.

There were six foreseen departures from the satisfaction of the safety criterion. The first four of these (1–4 below) do not seem to be readily remediable. They are clearly rare. A quantitative analysis might show that they are so rare that they can be ignored. However, data that would enable such a quantitative analysis were not collected for this study. They are listed below because they cast light on the inherent safety problems of this system, which may well also be present in most or all similar systems.

These foreseen hazards are as follows:

1. It is not usually a catastrophe if some malfunction of a vehicle in a platoon causes vehicles in the platoon to crush together. The collisions occur at low relative speeds, and the dividers ensure that the vehicles come to rest without striking anything else. In a low-speed collision, also, the lateral control systems should keep the vehicles on lane. If such an accident occurs at a gate, however, some wrecked vehicles, perhaps with occupants, could trespass onto another lane and be struck by another platoon. There have to be gates in the dividers because all vehicles must change lanes. This class of accident seems unavoidable; in fact, it will be rare. It requires (a) that a low-speed collision occur; (b) that it occur in a small space within a gate; and (c) that it damage lateral control systems sufficiently to cause them to malfunction. If the lateral control systems have some robustness, malfunction following a low-speed collision should be rare. But until the detailed design of the system is known one cannot say how rare.

A parallel situation arises if debris from an accident between manually controlled vehicles on parallel lanes, or a dropped load, protrudes onto the automated lanes. Alternatively, the debris from an accident on manual lanes can be so massive that it breaches the divider. In these cases there would have been an accident in the left-hand lanes even if there had been no automation; the accident is in no sense caused by faulty automation. However, the greater density of traffic in the automated lanes can mean that the number of casualties is greater than it would have been.

2. During the merge-and-split maneuvers, Hazard 1 (see above) is necessarily violated for a brief period. Platoons in automated lanes are separated by less than platoon spacing.

3. An object dropping from above onto the automated lanes presents a situation in which there would have been an accident in the absence of automation; however where there is increased density of traffic on the automated lanes there is an increase in the number of casualties.

4. If a manually controlled vehicle (illegally) enters the automated lanes, unpredictable behavior by the driver can give rise to hazards. It is a weakness of any design in which the intelligence is concentrated in the vehicles that a "rogue" cannot be tracked. In this case, the fact that the rogue will not respond to messages means that it is not detectable by sending messages. Hazards can then arise in the change-lane maneuver. Further, if the forward sensor requires a responder, the responder may be absent from a rogue. In such a case the rogue may not be detected at full range by the forward probe. There are no easy means of predicting the frequency of this occurrence or of its propensity to result in accidents. Much will depend on the probability that rogue drivers can be apprehended.

5. A fifth foreseen hazard is a design error in the original specification of Hsu et al. (*8*). Because this design is based on the work of Hsu et al. it was not thought right to remove it, even though it meant that the safety criterion would be violated. The change-lane maneuver permits a vehicle to enter just ahead of a platoon, which is hazardous because, if the entering vehicle strikes the gate, a high-speed collision results. This is readily remedied by requiring that the platoon in the receiving lane drop back a full platoon spacing.

6. The merge-and-split functions involve situations in which Hazard 1 arises. However, the merge operation is not necessary, provided the restriction that only free agents can change lanes is relaxed. If a vehicle always changes lane either into a large gap or immediately to the rear of a platoon, there is no hazard in that lane. If it leaves a lane from the center of a platoon, there is no hazard in that lane. However, because this work was based on that of Hsu et al. (*8*), these changes were not made in this work.

The fault tree analysis detected all these errors.

Some other errors were also detected, which were real errors. The fact that the fault tree picked them up demonstrates that the method is effective. It also demonstrates that mistakes in design are easily made and that the verification and validation process is indeed necessary to ensure that design meets safety criteria.

The following are design errors that were not foreseen but were detected by the fault tree:

1. If the fault of "loss of the forward sensor" is present at entry, it will not be detected immediately, and hazards can follow. The fault can be corrected by including a check on the forward sensor as part of the entry procedure.
2. If a vehicle is moving slowly and another wishes to change into the lane in which it is moving, the slow mover may not respond to the initial message (request_change_lane) because it is too far away. Nevertheless there is a possibility that the "changer" will catch up with it and change lanes too closely. As the design stands, a lane containing a vehicle that cannot communicate laterally is made a no-entry lane. Lanes with vehicles that are slow moving should also be treated in this manner.
3. If two vehicles each wish to change lanes into the same lane, but from opposite sides, and their speeds are ill matched, a hazard can arise. When the messages request_change_lane were sent, the vehicles were too far apart to need to reply; however, it is possible that at the time of change they are too close.
4. On exit, if a vehicle has a fault in its forward sensor and there is a manually controlled vehicle ahead of it, the separation of the two is unknown. This violates Hazard 2. Any danger can be avoided by causing such a vehicle to leave at a very low speed and to maintain such a low speed until manual control is resumed.

SUMMARY

A particular design to demonstrate the techniques of complete specification and fault tree analysis has been examined. These techniques have been successful here and in another example (*2*). It seems reasonable to conclude that they are suitable methods for examining the safety of all automated highway design concepts.

In the system considered here, lane changes were made only by free agents, and in the process of becoming a free agent, a split maneuver involves a situation in which Hazard 1 arises. This was done because the ideas of Hsu et al. (*8*) were being followed. An alternative is to require that lane changes take place by a vehicle's leaving a platoon from any position and either joining another at the rear or entering a large gap.

The particular design tested is not being suggested as a contender for construction. It is therefore not necessary to reiterate the cycle and attempt to correct the errors. Indeed, the detection of the errors is evidence of the effectiveness of fault tree analysis. However, correction of the unforeseen faults is readily done, as has been shown. It is reasonable to hope that the foreseen ones are rare enough to be ignored, but this remains to be demonstrated.

The following conclusions are made:

1. A specification has been made of an automated freeway. A set of hazards and a safety criterion also have been specified. A number of foreseen circumstances in which the safety criterion is not met seem likely to be rare enough to be accepted.
2. Fault tree analysis was practical. No branch of the tree had more than four elements. Faults were detected, but it is reasonably clear that they can be corrected.
3. It is therefore possible to construct an automated freeway that meets these safety criteria.

SAFETY OF ALTERNATIVE CONCEPTS

The relative merits, from a safety viewpoint, of alternative automated-freeway design concepts will now be considered. The consideration is based on the experience gained in the work reported here plus that of earlier parallel study (*3,4*). That was a study involving a specification and fault tree analysis of a system with a single automated lane and intelligence concentrated in the infrastructure.

The infrastructure-based design (*2,3*) relied heavily, perhaps too heavily, on the integrity of communications. This case is an extreme in the infrastructure basis of intelligence. The control data are passed from vehicle to vehicle in a platoon via the roadside. Any interruption to communication lasting several seconds could produce large disturbances in platoon motion, although the delay would have to be long and other independent disturbances would have to occur also before a hazard arose. How long and how severe are not yet established. But certainly, the more intelligence is concentrated in the infrastructure, the more important the integrity of road-vehicle communication becomes to safety.

Some vehicle-based intelligence is required to keep a vehicle hazard free if communications (vehicle-vehicle or vehicle-road) are less than totally reliable. However, this is not the only need. The system discussed here relies heavily on the performance of the forward sensor. This device, as specified, can be relied on to detect the vehicle ahead of one in the same lane at distances of more than a platoon spacing—perhaps up to 200 m. Roads curve, both vertically and horizontally. At their edges, roads have obstacles to vision. Whether clear sight lines can be guaranteed at this range (the sensor is unlikely to be above the level of the hood) can be doubted. Even more doubtful is the ability to guarantee that an image is that of a vehicle in the same lane. The forward sensor, as specified in this system concept, is not obviously readily implemented.

Alternatives to the use of a forward sensor are possible, In the system discussed in this report we control a vehicle that has lost its forward sensor by using vehicle-vehicle communication. In an earlier report (*2*) the same effect was achieved by the use of vehicle position detectors and a good deal of roadside logic. However, the discussion has turned full circle: it started with an examination of alternatives to totally reliable communications.

There is not necessarily an impasse. A balance of integrity between sensors and communications and a balance of intelligence between the roadside and the vehicle seem likely to offer opportunities to resolve the problem. However, investigation of the safety of such systems would rely heavily on the ability to predict reliability in quantitative terms. One would also need to be able to predict quantitatively the consequences of the loss of reliability, which would involve better understanding of the frequency with which critical configurations occurred in practical operation.

It is concluded, extending the result here, that it is possible to design an automated freeway that meets a required standard of safety. The immediate result here, however, is that it is possible to design an automated freeway that satisfies the safety criteria discussed in this paper.

ACKNOWLEDGMENTS

This work was performed as part of the program of California PATH of the University of California, in cooperation with California's Business, Transportation and Housing Agency and Department of Transportation and FHWA and NHTSA, U.S. Department of Transportation.

REFERENCES

1. A. Hitchcock. Intelligent Vehicle Highway System Safety: Problems of Requirement Specification and Hazard Analysis. In *Transportation Research Record 1318*, TRB, National Research Council, Washington, D.C., 1992, pp. 98–103.
2. A. Hitchcock. *Methods of Analysis of IVHS Safety. Part II: Main Report.* PATH Research Report UCB-ITS-PRR-92-14. University of California, Berkeley, 1992.
3. A. Hitchcock. *A Specification of an Automated Freeway.* PATH Research Report UCB-ITS-PRR-91-13. University of California, Berkeley, 1991.
4. A. Hitchcock. *Fault Tree Analysis of an Automated Freeway.* PATH Research Report UCB-ITS-PRR-91-14. University of California, Berkeley, 1991.
5. A. Hitchcock. *A Specification of an Automated Freeway with Vehicle-Borne Intelligence.* PATH Research Report UCB-ITS-PRR-92-16. University of California, Berkeley, 1992.
6. A. Hitchcock. *Fault Tree Analysis of an Automated Freeway with Vehicle-Borne Intelligence.* PATH Research Report UCB-ITS-PRR-92-15. University of California, Berkeley, 1992.
7. S. E. Shladover. *Operation of Automated Guideway Transit Vehicles in Dynamically Reconfigured Platoons.* Urban Mass Transit Administration Reports UMTA-MA-06-0085-79-1, UMTA-MA-06-0085-79-2, and UMTA-MA-06-0085-79-3, Springfield, Va., 1979.
8. A. Hsu, F. Eskafi, S. Sachs, and P. Varaiya. *The Design of Platoon Maneuver Protocols for IVHS.* PATH Research Report UCB-ITS-PRR-91-6. University of California, Berkeley, 1991.
9. P. Varaiya and S. E. Shladover. *Sketch of an IVHS System Architecture.* PATH Research Report UCB-ITS-PRR-91-3. University of California, Berkeley, 1991.

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Investigations into Achievable Capacities and Stream Stability with Coordinated Intelligent Vehicles

B. S. Y. Rao, P. Varaiya, and F. Eskafi

Determination of the achievable capacity of an intelligent vehicle highway system (IVHS) was undertaken. The IVHS architecture proposed by earlier researchers in which vehicles coordinate their behavior to form short-headway platoons on an automated lane was examined. Although many estimates of potential capacity of such an IVHS have been derived, they all assume that steady-state flow already has been reached. In such an IVHS, entrance and egress of vehicles are the primary causes of traffic stream disturbance, ultimately dictating the flow rates that can be sustained reliably. Therefore the focus of the investigation was on entrance and egress strategies. A detailed simulator, SmartPath, models the passage of individual intelligent vehicles along the highway and was used to examine transient behavior of the traffic stream under various conditions. Three different strategies were examined for allowing vehicles to enter and leave automated lanes and measuring the corresponding maximum flow rates that are attained. The length of time it takes vehicles to enter an automated lane and therefore to build up high flows was also measured. Finally the effects on flow of vehicles leaving an automated lane were measured, thus indicating that this would be the major source of traffic stream disturbance. It was concluded that, although maximum theoretical capacity cannot be attained, through prudent design of entrance and egress strategies, extremely high throughput can be sustained.

Many claims have been made about the potential capacity benefits of intelligent vehicle highway systems (IVHS). Close formation platooning (*1–3*) appears to offer the greatest potential of all systems proposed for capacity increase through automation, but no reliable measure of attainable capacity is available. Most calculations on potential capacity increases rely on a simple kinematic calculation assuming some maximum platoon size (N), inter- and intraplatoon spacing (d and Δ), vehicle length (l), and average speed (v). For example, to calculate flow in vehicles per hour per lane we can use

$$\text{Flow} = \frac{3600vN}{N(l + d) - d + \Delta} \qquad (1)$$

Because this assumes that all platoons are at maximum size, the average platoon size $\overline{N}$ is sometimes substituted for N (*4*). Such calculations implicitly assume that high levels of flow can be built up arbitrarily on a lane without inducing instability of the traffic stream or affecting flow in other lanes in the process. Furthermore these forms of analysis give no information of the tolerance of the traffic stream at such high flows to anticipated regularly occurring disturbances such as those caused by entrance and egress of vehicles from automated lanes.

California PATH, Institute of Transportation Studies, University of California, Berkeley, Calif. 94720.

The claim of this paper is that the major source of capacity limitation in such an IVHS will arise from behavior of the traffic stream as vehicles attempt to join and leave the automated lanes (*5*). To determine the attainable capacity of a close-formation platoon IVHS it is necessary to examine in detail how high flows will be built up on a lane in practice and also how such high flows will withstand the disturbances mentioned above. To this end we employ a detailed simulator, SmartPath (*6*), which models interactions between intelligent vehicles operating within the architecture described earlier (*3*). This level of detail allows one to derive quantitatively the effects of entrance and egress of vehicles to and from an automated lane and to test various strategies for achieving lane changes, thereby allowing the optimization of this key maneuver.

It is concluded that whereas entrance and egress of vehicles cause stream instabilities, it is still possible to attain substantial increases in capacity through platooning, although not by quite as much as may be predicted through simple kinematic calculations.

THE SIMULATOR

SmartPath simulates an automated highway system (AHS). Details of various aspects of the particular AHS that is simulated are available in several Partners for Advanced Transit and Highways (PATH) reports (*3,7–9*). SmartPath is a microsimulation: the system elements and control policies are each individually modeled. Several elements and policies are specified parametrically, allowing the user to change these parameters and thereby investigate any performance variations that may be caused.

SmartPath is built using the process-oriented simulation environment Csim (*10*). Each vehicle on the AHS is modeled as a separate process under Csim, facilitating an accurate representation of the parallelism of each vehicle's operation. The simulation models the behavior of vehicles from the moment they attempt to enter the AHS from an on-ramp until the time they leave at an off-ramp. Vehicles attempt to reach their (individually) prespecified lane, and if that lane is an automated one, they attempt to form platoons. SmartPath models the communication between vehicles that is necessary to achieve these goals (*8*).

The simulator updates information about each vehicle at a rate of 0.1 sec, simulation time. This rate of update allows SmartPath to model realistically the dynamic behavior of vehicles as they follow various control policies, adjusting acceleration and steering to perform various maneuvers.

Maneuvers

SmartPath allows vehicles to perform only three basic maneuvers, the communication protocols and detailed description of which may be found in earlier work (*8*). The three maneuvers are merge, split, and change lane.

- Merging allows a single vehicle or platoon in an automated lane to join the platoon in front and become one platoon. This can occur only when the two parties are close enough to communicate and initiate the process and if the new platoon will be smaller than the maximum allowed platoon size.
- Splitting breaks a platoon in an automated lane into two separate, autonomous platoons. The new platoon at the rear must decelerate until it is a safe distance behind the one in front.
- Changing lanes allows vehicles to change safely from their lane to the adjacent one. A lane change can occur only when a sufficiently large gap is present in the destination lane. If such a gap is not present, the vehicle that wishes to change and the obstructing vehicle or platoon decide between themselves how to create a gap by decelerating one or the other party or by splitting a platoon to make the gap. Only single vehicles are allowed to perform this maneuver. A single vehicle platoon is called a free agent.

Using these three maneuvers as basic building blocks, larger and more complex operations, such as changing lanes to join a platoon or exiting the automated lane, are achieved. Note that, as specified previously (*3*), only one maneuver per platoon is permitted at any time. These maneuvers are implemented using detailed control algorithms proposed for automated vehicles (*11,12*). SmartPath therefore accurately simulates the behavior of automated vehicles that operate using these algorithms.

Manually Controlled Vehicles

SmartPath currently has no satisfactory way of modeling the behavior of a completely manually driven vehicle. One of the many car following models proposed (*13–15*) could be implemented to overcome this deficiency. However, to use SmartPath to study situations in which coordinated automated vehicles interact with other types of vehicles we assume that in our AHS all vehicles are equipped with at least the technology to perform autonomous intelligent cruise control (AICC), which is deployed at all times on the highway, except when coordinated operation is selected by the driver. (We define AICC technology as a means of implementing the longitudinal control of a vehicle on the basis of the measurement of its speed and distance relative to the vehicle in front of it, together with measurements of its own state.)

Although this appears to be a major assumption it is arguable that the imminent advent of commercially available AICC systems will ensure that long before the first coordinated vehicles are ever sold a large proportion of vehicles will be AICC equipped. Essentially AICC is the tracking law as stated above that allows us to model "manually controlled vehicles" within SmartPath as platoons of maximum possible size so that the tracking law will always apply.

Example Output

The output of SmartPath can be displayed as a time-distance diagram for each lane. Figure 1 shows a detail from such a diagram for an automated lane. In the diagram shown some platoons have already been formed and vehicles entering the lane attempt to join those platoons. The diagram shows how platoons adjust their speeds to maintain safe headways as the situation around them changes.

EXPERIMENTAL FRAMEWORK

The general framework used is shown in Figure 2. We simulate a two-lane highway with several on- and off-ramps. We assume that the inside lane is the automated lane and that the adjacent lane is a transition lane in which traffic from an on-ramp is attempting to merge into the automated lane, or traffic from the automated lane is attempting to leave at the next off-ramp. Currently, therefore, we assume no permanent flow in this lane (see the section on shortcomings). We define

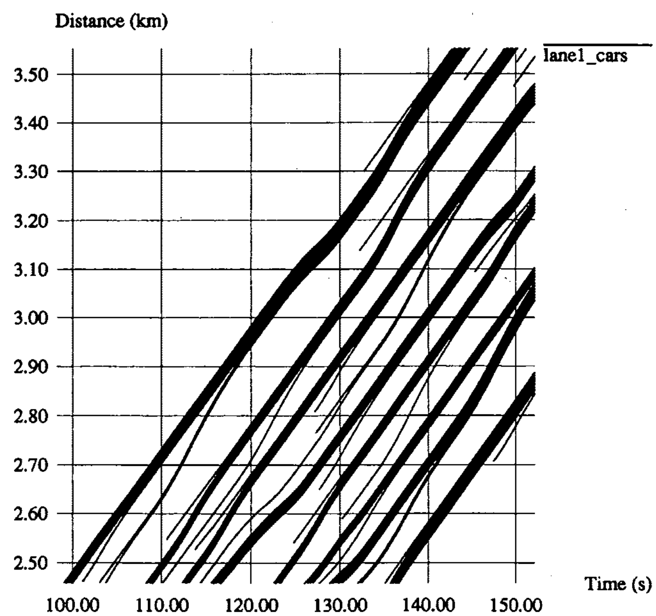


FIGURE 1 Portion of a time-distance diagram of the output of SmartPath for an automated lane. Many platoons already have formed by this stage, and individual vehicles that enter the lane attempt to merge with these platoons. Platoons adjust their speeds to maintain safe headways at all times.

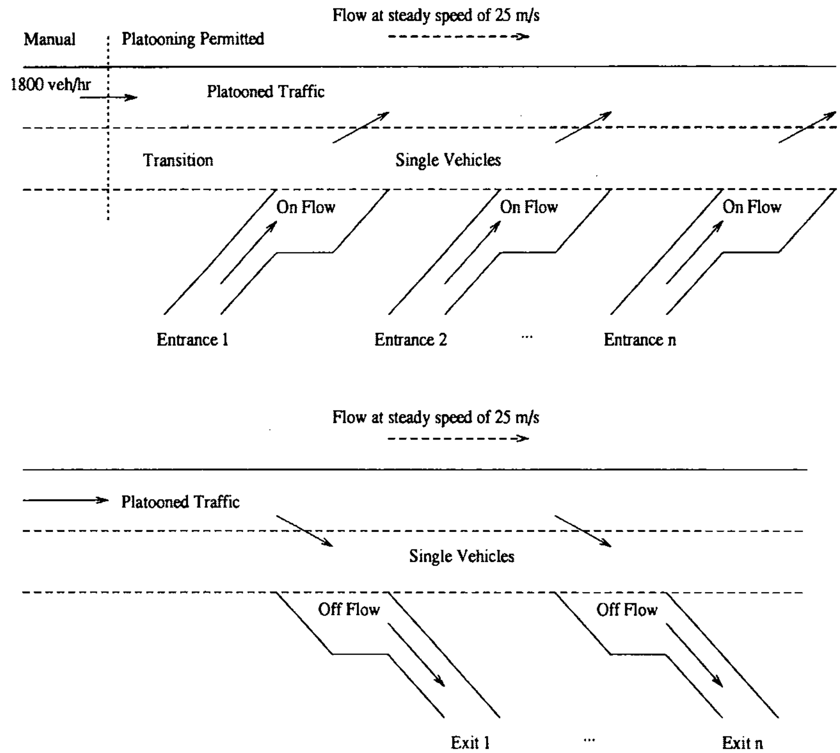


FIGURE 2 Diagram of the scenario modeled.

permanent flow in the transition lane as being those vehicles that travel in that lane to their destination without any desire to use the automated lane.

Initially a stream of vehicles at 1,800 vehicles/hr is on the inside lane, traveling at 25 m/sec under manual control. The vehicles are spaced randomly with headways drawn from a distribution of headways determined through measurement of traffic on highways (*16*). All vehicles are equipped with automation technology. At a certain point (the 0-km mark on all time-distance diagrams in this paper) all vehicles on the inside lane are able to use their automation facilities, and they begin to form platoons according to the methodology described above.

After some distance the traffic stream will be at steady state, and after this distance we place the first on-ramp. We define traffic in a section of highway to be in steady state when

1. All vehicles in that section are traveling at the desired constant speed,
2. All vehicles are in platoons (possibly of size 1) whose lead vehicle either cannot detect the platoon in front of it or cannot platoon with it because of maximum platoon size restrictions, and
3. No vehicles in that section are attempting to begin a maneuver.

Each subsequent on-ramp is placed such that traffic that entered from the previous on-ramp has merged into the automated lane and this lane is once more at steady state (see Figure 3). Off-ramps are placed similarly.

Using this framework we wish to determine the following:

- What is the maximum attainable flow in the automated lane?
- What is the maximum sustainable on-ramp flow rate?
- What is the distance that it takes vehicles entering at an on-ramp to merge into the inside lane?
- What is the effect on stream stability of vehicles entering and leaving?
- What is the maximum rate of egress of vehicles that can be supported?
- How long does it take vehicles to leave the automated lane?

Because all of these are, to some extent, random variables that are dependent on the initial headways of manual traffic on the inside lane and on how exactly platoons are formed, we obtain results after Monte Carlo simulations to attempt to average out the effects of the random behavior. The simulation is run 100 times with different starting headways (all drawn from the same distribution) and the cumulative results are analyzed. To allow results to be obtained within a reasonable time frame we generate a 40-sec section of traffic and monitor the progress of this "sample" as it proceeds along the highway. Flow rates are therefore scaled according to the subsequent time spread of this sample given that it started as a 40-sec sample. For example, if after some traffic has been added to obtain a total input flow of ϕ it is found that the sample now occupies 50 sec (because of deceleration of some vehicles to make room for others), we must calculate the effective flow rate as $(40/50)\phi = 0.8\phi$. This applies in reverse if the sample is found to compress.

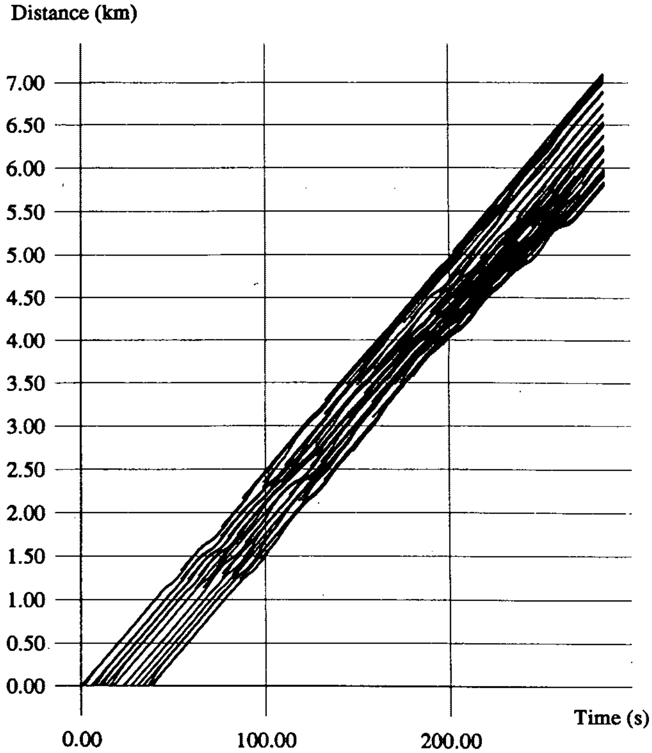


FIGURE 3 Time-distance diagram for the automated lane in an AHS using entrance-egress strategies described in Case 1.

Assumptions

We now state the assumptions on the behavior of the automated vehicles and choose values for critical parameters such as interplatoon distance.

Behavior of Vehicles

We assume that platooning is possible only in the automated lane and that vehicles are able to communicate occasionally with each other to perform maneuvers and also frequently with each other when entrained as platoons. We assume that drivers are able to activate their automation equipment at some point either when their lane becomes the automated lane or when they are in the transition lane after entering from an on-ramp. We assume that each driver does this at the earliest opportunity to ensure that vehicles enter the automated lane as soon as possible. Once under automatic control we assume that the vehicle will enter the automated lane initially as a free agent, although there may be other vehicles nearby (but a safe distance away). We assume that a vehicle that is a free agent or a leader of a platoon will follow a control law that will keep the vehicle at the defined target speed and always at a safe distance (Δ) from the vehicle in front.

We also assume that vehicles in the transition lane follow this control law, consistent with the previous assumption that all drivers enable their automation equipment on entering the transition lane.

We assume that vehicles travel on the automated lane until close to their exit. They then initiate a maneuver to leave the lane; once they are in the transition lane drivers resume control of the vehicle, allowing them to exit.

Range of Detection

A fundamental parameter is D, the range of detection of a vehicle. Although line-of-sight devices with ranges of up to 150 m have been reported (*17*) this figure may change radically in adverse lighting conditions, poor weather, and when rounding a bend. We assume that the simulated highway segment is largely straight and that the weather is fair; thus, D = 70 m. We also assume that this sensor is sampled at 10 Hz, giving a possible maximum sensing delay of $\tau_s = 0.1$ sec.

Rates of Acceleration and Deceleration

Acceleration and deceleration rates vary enormously with road conditions, tire conditions, and weather. It has been suggested that deceleration rates of only 0.25 g can be attained under the worst cases (*18*), but recent field tests by PATH researchers at San Diego have shown that, under optimal conditions, deceleration rates of 0.9 g can be achieved for emergency braking. For comfort we limit the maximum braking and acceleration to $-0.2\,g$ and $0.2\,g$, respectively, but during emergencies we assume the maximum deceleration $a_{\max_d} = -0.5\,g$ can be achieved.

Actuator Delays

We assume that the actuator delay τ_a is 0.2 sec for full braking to be effected. This value is based on experiments with actuators to be used in PATH test vehicles.

Speed

The speed depends on D, the maximum deceleration, $a_{\max_d}$ and a reaction time $\tau_r = \tau_a + \tau_s$. We set the speed such that single vehicles (or platoons) can stop before colliding with any vehicle decelerating in front of it. Because vehicles can only see a distance D ahead, we conjecture that the worst case for a vehicle (or platoon) is when an object that is already stationary is detected at the limit of its detection capability. Therefore

$$v = \sqrt{a_{\max_d}^2 \tau_r^2 + 2a_{\max_d} D} \qquad (2)$$

which with figures given yields 24 m/sec $< v <$ 28 m/sec. We set v = 25 m/sec as a convenient value for our simulations.

Interplatoon Distance

Because the speed was chosen to enable safe stopping on encountering a stationary object, interplatoon spacing (Δ) must serve to prevent collision with a platoon that is always within range of detection. If the platoon in front should brake suddenly and the platoon behind is able to apply similar braking power after some reaction period (τ_r), we can say

$$\Delta \geq \frac{1}{2} a_{\max_d} \tau_r^2 + \tau_r v \tag{3}$$

which, given our assumptions, gives $\Delta \approx 8$ m. This is infeasible for two reasons:

- Without some method for checking that all vehicles entering the AHS satisfy criteria concerning minimum levels of performance of equipment it cannot be assumed realistically that every platoon has similar braking characteristics. It may be that the trailing platoon will not be able to decelerate at the same rate as the one in front because of worn brakes or poor tire quality, necessitating a significant increase in Δ (*18*).
- Even if there existed a method for ensuring that the above assumption could be made, there is (to our knowledge) no lead vehicle control law that can maintain such spacings without requiring a huge control effort or frequent interplatoon communication (not supported in our architecture).

Therefore we increase Δ to 30 m, which is more reasonable from a control standpoint and is also safe.

Intraplatoon Distance

This distance parameter (d) is the spacing between the front of a car and the rear of the preceding car in a platoon and is set to 1 m. This value of d will lead to a (low relative velocity) collision between two consecutive vehicles in a platoon if their communication system fails and a rapid deceleration is called for before this failure is detected. The frequent communication system is checked several times a second and any vehicle failing to receive information from its platoon leader could simply decelerate gently, preventing any collisions in the remainder of the platoon, until it is at rest. In other instances the communicated information should keep platoon partners from hitting each other (*12,19*) even in the event of sudden braking.

Maximum Platoon Size

Platoon size (N) is not an independent variable but rather, because of the adaptive nature of this system, varies with the demand flow. However, it is currently thought that platoons larger than 20 are difficult to control and, in fact, large platoons make egress more difficult, so we set $N_{\max} = 20$.

Shortcomings in the Simulation Methodology

One area in which SmartPath is currently deficient is the control law for the lead vehicle of a platoon. (To our knowledge no clear control law for lead vehicles in platoons has been proposed in the published literature. Merely maintaining a constant speed is not sufficient to cope with all eventualities.) The law used calculates the deceleration a vehicle must apply at each instant on the basis simply of the distance and relative velocity of the vehicle in front such that a collision is always avoided (*6*). The same law is also used for vehicles in the transition lane, and its crude nature leads to situations of traffic stream instability that we believe can be avoided through prudent design of this law. It is therefore not currently possible to simulate any permanent manual flow in the transition lane, although there is work in progress to develop a more satisfactory lead vehicle control law.

THREE CASE STUDIES

We examine three different policies governing entrance to and egress from the automated lane, all implementable using the basic merge, split, and change-lane maneuvers.

Case 1: Basic Model

In this study the entrance and egress policies are as follows. Vehicles entering the automated lane from a transition lane do so immediately if no vehicles are within the safe distance. If a vehicle in the transition lane is adjacent to a platoon it will enter in front of it if the vehicle is toward the front third of the platoon, behind it if the vehicle is toward the rear third of the platoon, or the middle of the platoon otherwise. In all cases the whole platoon, split platoon, or vehicle must decelerate to the safe interplatoon distance before the lane change occurs. Merging may then take place as a separate maneuver afterward. Vehicles that wish to leave the automated lane do so from the rear, front, or middle of the platoon, but all of these maneuvers entail a deceleration to safe separation before the lane change.

In the scenario simulated for this case, vehicles enter the automated lane (which starts at 0 km) and begin to platoon together. On-ramps at 1, 2, 3, and 4 km are supplying vehicles at a regular metered rate of one every 2 sec (1,800 vehicles/hr), all of which attempt to merge into the automated lane, thereby theoretically achieving flows on the automated lane of 3,600, 5,400, 7,200, and 9,000 vehicles/hr. There is an exit at 7 km, and vehicles attempt to leave the automated lane from about 5.5 km at a rate of 900 vehicles/hr.

From Figure 3 we can see that although high flows can be attained, rider comfort would be low because each platoon decelerates and accelerates many times in reaction to vehicles in front. The primary culprit of this behavior is the vehicles entering the automated lane before joining a platoon. In this model, such vehicles must decelerate to Δ behind the platoon before joining the lane and then merging with the platoon. This causes a large change in headway to the platoon behind, which must react, although the change in headway is merely transient.

From Figure 4 we see that high on-ramp flows can be supported. It can be argued that 1,800 vehicles/hr is close to the maximum that an on-ramp can be expected to supply, given that the lane capacity of a highway now is around that value.

Vehicles leaving the automated lane cause large stream disturbances because splitting to allow vehicles out results in the back portion of the platoon decelerating until it is Δ behind

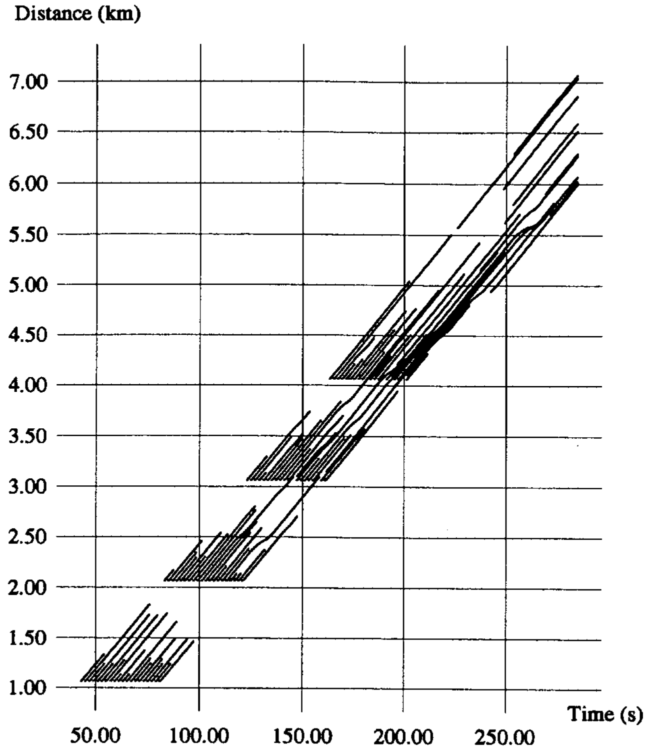


FIGURE 4 Time-distance diagram for the transition lane using entrance-egress strategies described in Case 1. Vehicles leaving the automated lane are found in the transition lane after 7 km.

the vehicle that wants to leave. In this case an egress demand rate of only 900 vehicles/hr causes a 25 percent drop in flow (Figure 5).

Case 2: Improved Merge and Lane Change

To improve rider comfort, the strategy for joining the back of a platoon was altered such that a vehicle could now be considered part of the platoon it will eventually join from the moment its request for a merge into the automated lane is acknowledged. It knows it will join that platoon, so it enters the automated lane only $2d$ behind the platoon and then merges with it. This should smooth out traffic flow disturbances and also make this maneuver safer. We expect that when a vehicle's headway is less than Δ but not as low as d the vehicle is at greatest risk of a high relative velocity collision, and by minimizing the amount of time a vehicle spends in this "gray zone" we claim that the maneuver is rendered less dangerous.

In the scenario simulated for this case vehicles enter the automated lane (which starts at 0 km) and begin to platoon together. On-ramps at 1, 2, 3, and 4 km supply vehicles at a regular metered rate of one every 2 sec (1,800 vehicles/hr), all of which attempt to merge into the automated lane, thereby theoretically achieving flows on the automated lane of 3,600, 5,400, 7,200, and 9,000 vehicles/hr. There is an exit at 7 km, and vehicles attempt to leave the automated lane from about 6 km at a rate of 1,800 vehicles/hr.

Figure 6 shows that the strategy has the required effect for traffic joining the automated lane, and it can be seen that (near entrances) platoons deviate from optimal speed far less often and with less magnitude. High flows are again achieved, and Figure 7 shows that high on-ramp flows are maintained.

Once again it was found that egress of vehicles caused a large drop in flow (Figure 5). This is mainly because of the split maneuver that causes a large change in headway to the vehicle behind the splitting platoon. This problem is exacerbated when either more than one vehicle has to leave a platoon or a number of consecutive platoons split at about the same time, causing a cascading shock wave that travels upstream.

This is a potentially worrysome problem because it implies that vehicles must begin to try and leave the automated lane long before their exit and also coordinate their actions via a higher (link) level to minimize the chances of causing massive stream turbulence—the former being unappealing and the latter being difficult to implement.

Case 3: Sorting by Exit Positions

In this strategy we allow vehicles to leave a platoon only from the back. This strategy allows easy egress and causes minimum change in headway for the platoon behind because the exiting vehicle can pull out of the lane after decelerating only a few feet.

However, to allow this action to take place we must allow vehicles to join a platoon only if the rearmost vehicles will leave the lane before those in front. We therefore sort vehicles as they attempt to join platoons. This is easily accomplished within the framework of the architecture specified previously (*3*) and requires only a small extra burden on the lead vehicle. When a request for a merge with a platoon arrives it is accompanied by an exit number. The lead vehicle then passes this down the platoon until a vehicle replies that the exit number corresponds to an exit after its own. At that point in the platoon a split will occur and the new vehicle may join. This strategy will have several effects:

- Exiting the automated lane is made easier at the expense of greater turbulence in both the transition lane and the automated lane as vehicles enter.
- Vehicles will have to be fed into the highway at lower rates but at more numerous entrances to allow the transition lane traffic to sustain the flow disruptions.
- "Opportunistic" merging of platoons will not be possible now unless the exit number order is maintained. This type of merging occurs after a split to allow a vehicle out or, more importantly, if there is more than one automated lane or if two automated lanes merge at any time.

In the scenario simulated for this case there are on-ramps at 1, 2, 3, 4, 5, and 6 km that are supplying vehicles at a regular metered rate of one every 4 sec (900 vehicles/hr), all of which attempt to merge into the automated lane, thereby theoretically achieving flows on the automated lane of 2,700, 3,600, 4,500, 5,400, 6,300, and 7,200 vehicles/hr. There are 10 exits starting at 8 km and continuing at 1-km intervals thereafter. At 7.25 km, vehicles for the 8-km exit (about 1 in 10 of the vehicles on the automated lane) attempt to leave. Similarly at 8.25 and 9.25 km vehicles attempt to leave for their exits.

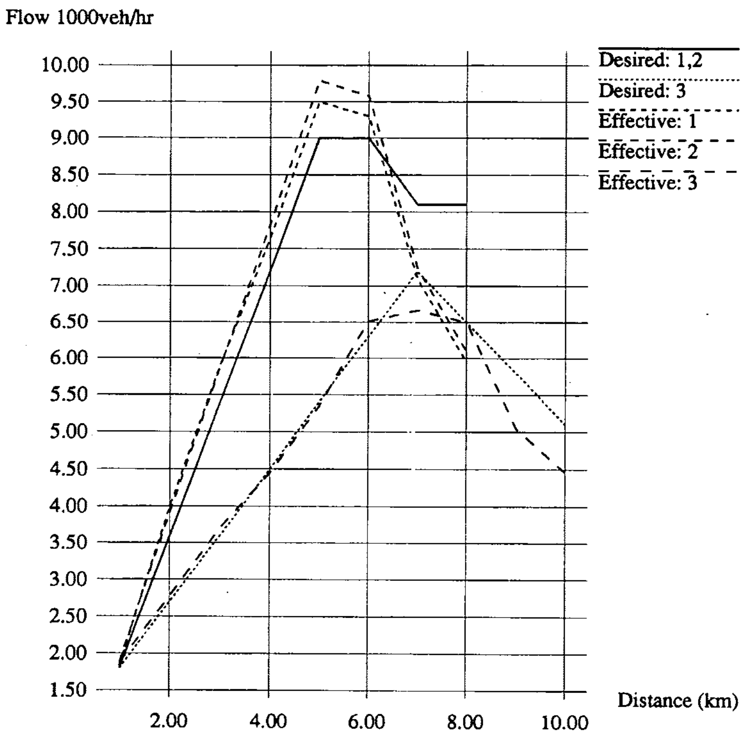


FIGURE 5 Effective flow actually found on the automated lane, measured at 1-km intervals, compared with input flow. Input flow is the existing flow on the lane plus any flow from an on-ramp that is trying to enter the lane or minus any flow that is attempting to leave the lane. Data for all three entering and exiting strategies are included. For Cases 1 and 2, the sustainable on-ramp flows are 1,800 vehicles/hr per ramp, but for Case 3 it is 900 vehicles/hr per ramp. Flows build up quickly, but in Cases 1 and 2 any exiting (at a rate of 900 vehicles/hr in these cases) leads to stream disturbances, which cause "spreadout" of the traffic sample, resulting in severe breakdown of flow. Case 3 shows how exiting of vehicles at this rate can be supported, although a lower maximum flow is initially attained.

Figure 8 shows that egress is indeed very smooth. In this simulation each vehicle was randomly allocated 1 of 10 possible exits so that at each exit approximately 1⁄10 of the maximum flow would try to leave (about 700 vehicles/hr). Note that the platoons that are formed are smaller but more numerous because random merging between platoons is now not possible.

A relatively high flow rate is achieved only after six low supply entrances (see Figure 9), and unlike the previous two cases in which effective flow exceeded input flow, in this case the flow begins to peak at about 6,500 vehicles/hr (Figure 5).

CONCLUSIONS

It is clear from our results that no matter what the potential capacity of an automated lane, it is the behavior of the traffic stream near entrances and exits that will ultimately determine access and egress rates, achievable flows, and rider comfort. In this paper we have discussed three possible strategies for allowing vehicles to enter and leave an automated lane. Appropriate design of these strategies can improve all of the above factors and also expedite the task of designing a suitable control law for lead vehicles by eliminating unwanted deviations in detected headways and relative speeds.

From our investigations we can state that platooning operating under the architecture and assumptions proposed previously (*3*) can achieve peak flows greater than 9,000 vehicles/hr rapidly and while maintaining safety and stream stability, but that these rates of flow cannot be sustained. Attempting to allow vehicles to leave causes significant stream disturbance, bringing the sustainable flow rate down to about 5,500 vehicles/hr. The main cause of stream disturbance at exit points is the current split maneuver. An attempt to eliminate the necessity for this maneuver near exits without requiring a change of protocols and basic maneuvers has been described but is seen to have several fundamental drawbacks.

ACKNOWLEDGMENTS

This work is supported by the California PATH program and the California Department of Transportation. The authors

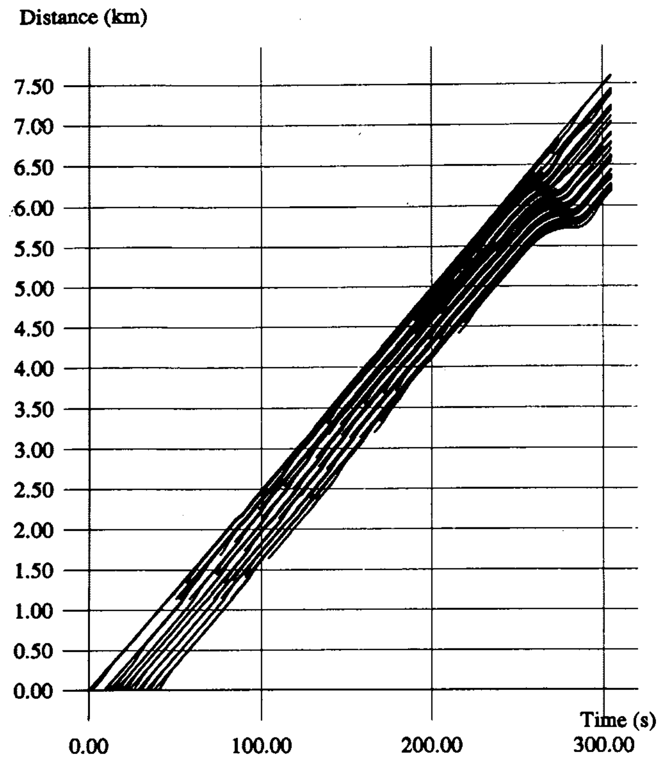


FIGURE 6 Time-distance diagram for the automated lane in an AHS using the entrance-egress strategies described in Case 2.

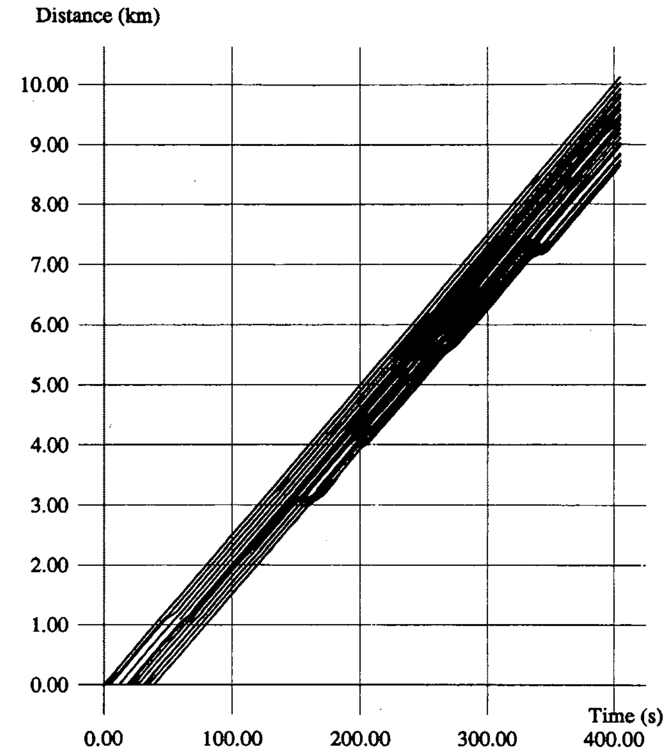


FIGURE 8 Time-distance diagram for the automated lane in an AHS using the strategy described in Case 3 in which platoons are formed according to the exit point of vehicles.

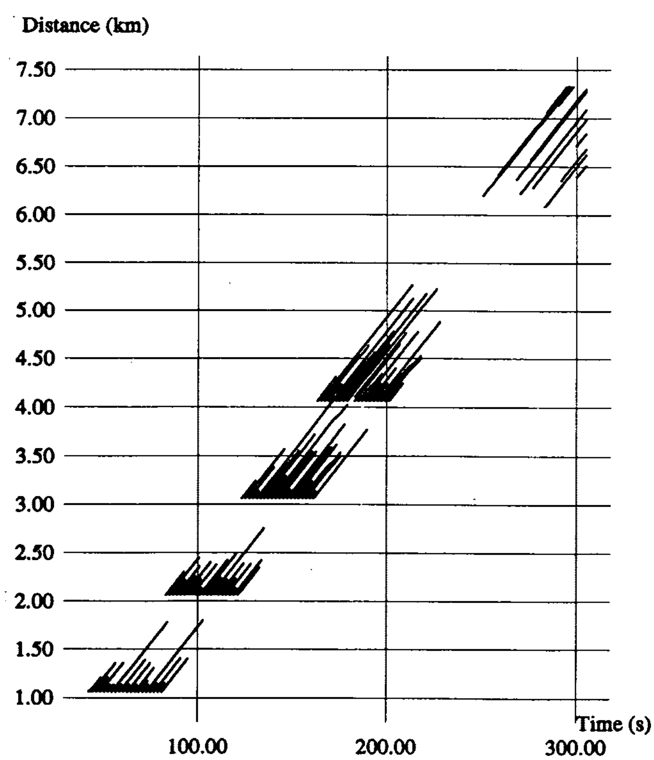


FIGURE 7 Time-distance diagram for the transition lane using entrance-egress strategies described in Case 2. Vehicles leaving the automated lane are found in the transition lane after 7 km.

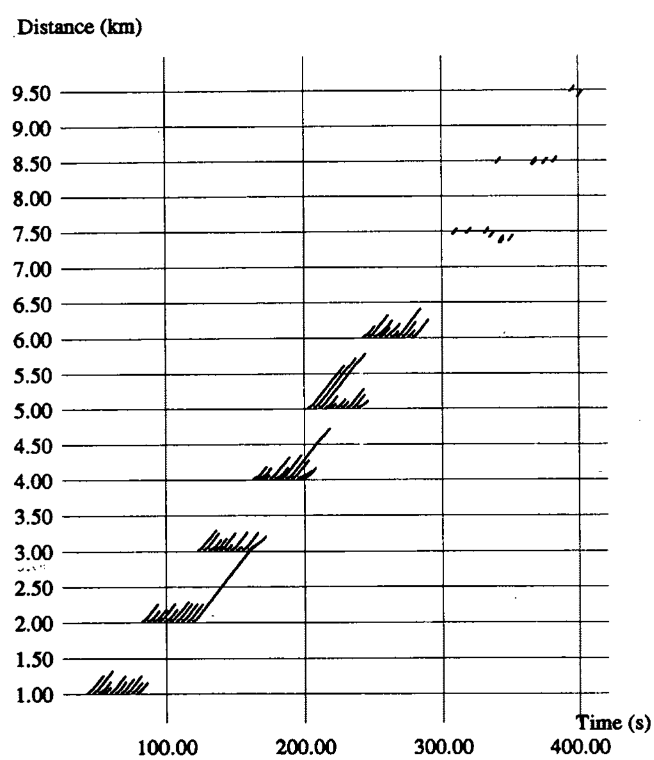


FIGURE 9 Time-distance diagram for the transition lane using strategy described in Case 3. Vehicles leaving the automated lane are found in the transition lane after 7 km.

would like to thank S. Sheikoleslam and S. Shladover for many useful comments and helpful guidelines.

REFERENCES

1. S. Shladover. Dynamic Entrainment of Autonomous Guided Transit Vehicles. *High Speed Ground Transportation Journal*, Vol. 12, No. 3, 1978, pp. 87–113.
2. S. Shladover. *Operation of Automated Guideway Transit Vehicles in Dynamically Reconfigured Trains and Platoons* (Vol. 2). Technical Report UMTA-MA-06-0085-79-3. UMTA, 1979.
3. P. Varaiya and S. Shladover. *Sketch of an IVHS Architecture.* Technical Report UCB-ITS-PRR-91-3. Institute of Transportation Studies, University of California, Berkeley, 1991.
4. U. Karaaslan, P. Varaiya, and J. Walrand. *Two Proposals to Improve Traffic Flow.* Technical Report Preprint. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 1990.
5. S. Shladover. Operation of Merge Junctions in a Dynamically Entrained Automated Guideway Transit System. *Transportation Research*, Vol. 14A, 1979, pp. 85–112.
6. F. Eskafi and P. Varaiya. *SmartPath: An Automated Highway System Simulator.* Technical Report 92-0708-1. California PATH, 1992.
7. S. Shladover, C. Desoer, J. Hedrick, M. Tomizuka, J. Walrand, W. Zhang, D. McMahon, H. Peng, S. Sheikoleslam, and N. McKeown. Automated Vehicle Control Developments in the PATH Program. *IEEE Transactions on Vehicular Technology*, Vol. 40, Feb. 1991, pp. 114–130.
8. A. Hsu, F. Eskafi, S. Sachs, and P. Varaiya. *The Design of Platoon Maneuver Protocols for IVHS.* Technical Report UBC-ITS-PRR-91-6. PATH, 1991.
9. P. Varaiya. *Smart Cars on Smart Roads: Problems of Control.* Technical Report PATH Tech Memo 91-5. Institute of Transportation Studies, University of California, Berkeley, 1991.
10. H. Schwetman. *CSIM Reference Manual* (Revision 13). Technical report. Microelectronics and Computer Technology Corporation, Austin, Tex., 1989.
11. J. Hedrick, D. McMahon, V. Narendran, and D. Swaroop. Longitudinal Vehicle Controller Design for an IVHS System. In *American Control Conference*, Boston, Mass., 1991, pp. 3107–3112.
12. S. Sheikoleslam and C. A. Desoer. Longitudinal Control of a Platoon of Vehicles with No Communication of Lead Vehicle. In *American Control Conference*, Vol. 3, 1991, pp. 3102–3107.
13. D. C. Gazis, R. Herman, and R. Potts. Car-Following Theory of Steady State Traffic Flow. *Operations Research*, Vol. 7, 1958, pp. 499–595.
14. H. J. Payne. Model of Freeway Traffic and Control. In *Mathematical Models of Public Systems, Simulation Council Proceedings*, 1971, pp. 51–61.
15. M. Papageorgiou. *Applications of Automated Control Concepts to Traffic Flow Modeling and Control.* Springer-Verlag, 1983.
16. A. D. May. *Traffic Flow Fundamentals.* Prentice Hall, 1990.
17. R. Muller and G. Nocker. Intelligent Cruise Control with Fuzzy Logic. In *IEEE Conference on Intelligent Vehicles*, Detroit, Mich., 1992, pp. 173–178.
18. A. Hitchcock. Coefficients of Friction and Automated Freeways. PATH MOU 19 Occasional Paper 1. May 1992.
19. S. Sheikoleslam and C. A. Desoer. *Longitudinal Control of a Platoon of Vehicles: Nonlinear Model.* Technical Report UCB-ITS-PRR-90-1. PATH, 1990.

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Flow Benefits of Autonomous Intelligent Cruise Control in Mixed Manual and Automated Traffic

B. S. Y. Rao and P. Varaiya

Autonomous intelligent cruise control (AICC) is a technology for driver convenience, increased safety, and smoother traffic flow. AICC also has been proposed for increasing traffic flow by allowing shorter intervehicle headways. Because an AICC-equipped vehicle operates using only information available from its own sensors, there is no requirement for communication and cooperation between vehicles. This format allows gradual market penetration of AICC systems, which makes the technology attractive from a systems implementation standpoint. The potential flow increases when only a proportion of vehicles on a highway are equipped with AICC were examined, and theoretical upper limits on flows as a function of pertinent variables were derived. Because of the limitations of the theoretical models, a simulator was used that models interactions between vehicles to give detailed information on achievable capacity and traffic stream stability. Results showed that AICC can lead to potentially large gains in capacity only if certain highly unrealistic assumptions hold. In reality, the capacity gains from AICC are likely to be small.

There is considerable interest in autonomous intelligent cruise control (AICC) technology within the advanced vehicle control system (AVCS) research community. For the purposes of this paper AICC technology is defined as a means of implementing the longitudinal control of a vehicle on the basis of the measurement of its speed and distance relative to the vehicle in front of it, together with measurements of its own state (*1–3*). Because these measurements can be obtained from sensors placed on the vehicle, the control law requires no cooperation from other vehicles—hence the term "autonomous." Consequently, AICC-equipped vehicles, in principle, can operate together with manually controlled vehicles. AICC technology thus provides an intermediate step toward the automated highway system (AHS) in which a vehicle's controller relies on information received from other vehicles.

Proponents of AICC make two sets of claims. The first set of claims concerns how well a single platoon of AICC-equipped vehicles follows a lead vehicle that undergoes a specified speed profile. The claims are about the stability of the platoon, the absence of the "slinky" effect, and the magnitudes of acceleration and deceleration induced by the control law. We shall not examine these claims here.

We examine here the second set of claims that concerns the increases in the capacity of a highway used by vehicles some or all of which are AICC equipped. The analysis supporting these claims is based on steady-state flow relations derived from the investigation of a single platoon. The validity of such an analysis rests, as we shall show, on a number of assumptions about the behavior of the lead vehicle of a platoon of AICC-equipped vehicles and the way in which (manually driven) vehicles merge into a lane containing platoons. We shall make these assumptions explicit.

We derive a theoretical upper bound on the capacity gain using AICC for various degrees of market penetration. We then describe results of a simulation in which the behavior of vehicles is modeled at a detailed level, showing how it has been applied to give information on achievable capacities and stream stability.

We conclude that AICC can offer modest improvements to lane capacity at low market penetration levels and probably has a beneficial, if slight, effect on stream stability. At higher levels of implementation, greater increases in capacity become harder to achieve because of problems of stream instability and limits on the rates at which vehicles can be fed into highways.

INITIAL MODEL FOR PREDICTING FLOW BENEFITS

We begin by modeling the process by which platoons are formed using AICC and attempt to derive theoretical upper bounds for capacity improvements for a range of degrees of implementation. "Platoons" in this case refers to groups of vehicles following each other at relatively close headways under AICC.

The basis formula to calculate flow in vehicles per hour is

$$\text{Flow} = \frac{3600v\overline{N}}{\overline{N}(l + d) - d + \overline{\Delta}} \qquad (1)$$

where

v = vehicle speed (m/sec);
l = length of a car (m);
d = spacing between cars in a platoon (m);
$\overline{\Delta}$ = average spacing between platoons (m); and
$\overline{N}$ = average size of a platoon:

$$\overline{N} = \sum_{1}^{\infty} Np(N)\, dN \qquad (2)$$

California PATH, Institute of Transportation Studies, University of California, Berkeley, Calif. 94720.

In an "adaptive" platooning scheme such as this, an increase in demand gives rise to a natural increase in platoon sizes and a decrease in interplatoon spacings. The relationship between $\overline{N}$ and $\overline{\Delta}$ is

$$\overline{\Delta} = \overline{N}(\overline{\Delta}_{man} - d) + d \quad (3)$$

The value $\overline{\Delta}_{man}$ is the average headway of manually controlled vehicles (before platooning starts).

It is clear that to derive the flow we must derive the platoon size distribution, $p(N)$.

In the ensuing analysis we denote any term pertaining to the vehicle in question with a subscript k and any term pertaining to the vehicle behind it with the subscript $k - 1$. The following terms are defined:

- $p(\text{man})$ = probability that a car is manually driven;
- $p(\text{aicc}) = 1 - p(\text{man})$ = probability that a car is AICC equipped;
- $p(\text{leader})$ = probability that a car is a leader of a platoon;
- $p(\text{follower}) = 1 - p(\text{leader})$ = probability that a car is in a platoon but not the leader;
- Δ_k = headway of Vehicle k before platooning begins;
- Δ_{max} = maximum distance a car can detect another car;
- $p(\Delta_k > \Delta_{max})$ = probability that Car k cannot detect the car ahead;
- $p(\Delta_k < \Delta_{max})$ = probability that Car k can detect the car ahead; and
- $p(\Delta)$ = probability distribution of vehicle headways before platooning begins.

SCENARIO

The modeled scenario is shown in Figure 1. There is a single manual/AICC lane (the inside lane in Figure 1), an adjacent "transition" lane, and several entrance ramps that allow vehicles to enter at a regular metered interval rate. We assume that a certain fraction of all vehicles have AICC capability. Vehicles are initially traveling at some flow ϕ along a single-lane highway under fully manual control. We assume that the flow in the manual part of the inside lane is restricted to 1,800 vehicles/hr, with headways drawn randomly from a distribution that models the way drivers behave on highways (*4*) (see Figure 2). At some fixed point along this highway those vehicles equipped with AICC are allowed to activate their systems. We assume that all vehicles equipped with AICC that can detect the vehicle in front of them in the manual section form a platoon with those vehicles once they have entered the automated section and that there is no limit on potential platoon size.

We assume that some vehicles from the on-ramps are attempting to enter the inside lane; because we are attempting to find an upper bound on flow in that lane, we assume that only AICC-equipped vehicles attempt to enter the inside lane from the transition lane.

We wish to derive the platoon size distributions for various input flows and for increasing proportions of vehicles equipped with AICC. This will allow us to calculate the flow on the highway once all vehicles have settled to steady state (i.e., all platoons have been formed).

Analysis

We begin by analyzing the formation of platoons in the inside lane as vehicles enter the AICC region (see Figure 3). To form a platoon of any size, the first requirement is for a leader. A leader is either a manual vehicle or an AICC vehicle that cannot detect the vehicle in front of it. Therefore

$$p_k(\text{leader}) = p_k(\text{man}) + p_k(\text{aicc} \wedge \Delta > \Delta_{max}) \quad (4)$$

For a platoon of size N there must be a leader (k) followed by a line of $N - 1$ AICC vehicles that are close enough to each other before platooning starts that they can each detect the preceding vehicle, followed finally by either a manual vehicle or an AICC vehicle that is too far to detect vehicle $k - (N - 1)$; in other words, another leader (see Figure 3). Therefore the probability of generating a platoon of size N from

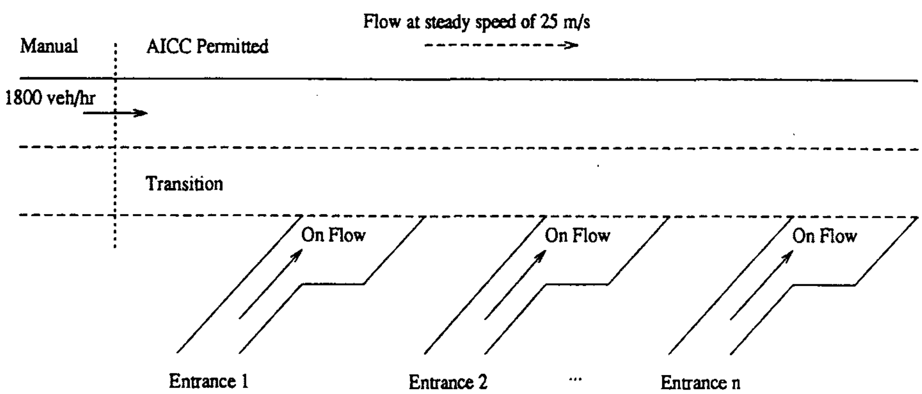


FIGURE 1 Diagram of the highway section modeled.

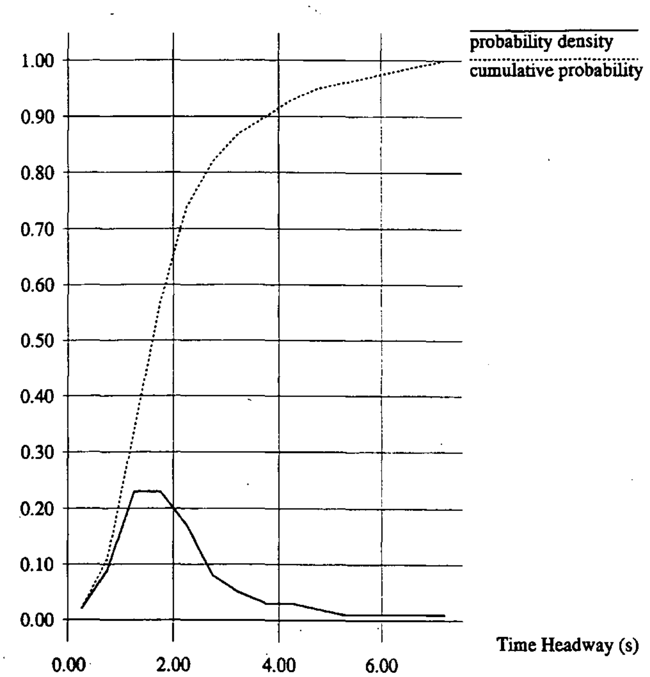


FIGURE 2 Maximum flow rate found on highways currently is about 1,800 vehicles per hour per lane. Drivers under such conditions exhibit a range of time headways. If a vehicle is traveling at 25 m/sec and can detect up to 60 m, it is clear that $p(\Delta) < \Delta_{max}) \approx 0.7$. This distribution is after May (*4*).

vehicle k is

$$p_k(N|k) = p_k(\text{leader}) \prod_{n=1}^{N} [p_{k-(n-1)} (\text{aicc} \wedge \Delta < \Delta_{max})] \, p_{k-N}(\text{leader}) \quad (5)$$

If we now assume that $p_k(\text{man})$ [and $p_k(\text{aicc}) = 1 - p_k(\text{man})$] is independent of $p_k(\Delta)$ and that $p_k(\text{man})$ and $p_k(\Delta)$ are independent of k, then we can rewrite as follows (dropping redundant k's):

$$p(\text{leader}) = p(\text{man}) + p(\text{aicc})p(\Delta > \Delta_{max})$$

$$p_k(N|k) = p(\text{leader})^2 \, [p(\text{aicc})p(\Delta < \Delta_{max})]^{N-1}$$

Because in a platoon of size N we could equally have started with any vehicle within that group of cars, we can say that our choice of vehicle k is a 1 in N chance for that particular group of cars. Therefore

$$p(N) = p_k(N|k)p(k) = p_k(N|k)(1/N)$$

$$p(N) = Np(\text{leader})^2 \, [p(\text{aicc})p(\Delta < \Delta_{max})]^{N-1} \quad (6)$$

This is a beta distribution, and it is clear that $p(N)$ is dependent on $p(\text{aicc})$, $p(\Delta)$, and Δ_{max}.

We then assume that after each on-ramp there are a certain number of AICC vehicles in the transition lane that wish to enter the inside lane. We model this situation as shown in Figure 4. Given any platoon in the inside lane, we assume that all AICC-equipped vehicles in the transition lane that are within a "catchment" region will eventually join that platoon. The size of the catchment region for a platoon of size N is $L(N)$ and is defined by the method by which vehicles join platoons. Because there is insufficient intervehicle spacing in a platoon for vehicles to merge into a platoon safely and it may not be possible for vehicles in the adjacent lanes to accelerate to join the front of a platoon, we assume that vehicles have to join the rear of platoons. Therefore, we can postulate a requirement for safe maneuvering distance δ in front of a platoon within which vehicles attempting to join the platoon in front cannot encroach. This defines the size of Catchment Region 3 to be

$$L(N) = \delta + N(l + d) - d + (\Delta - \delta)$$
$$= N(l + d) - d + \Delta$$

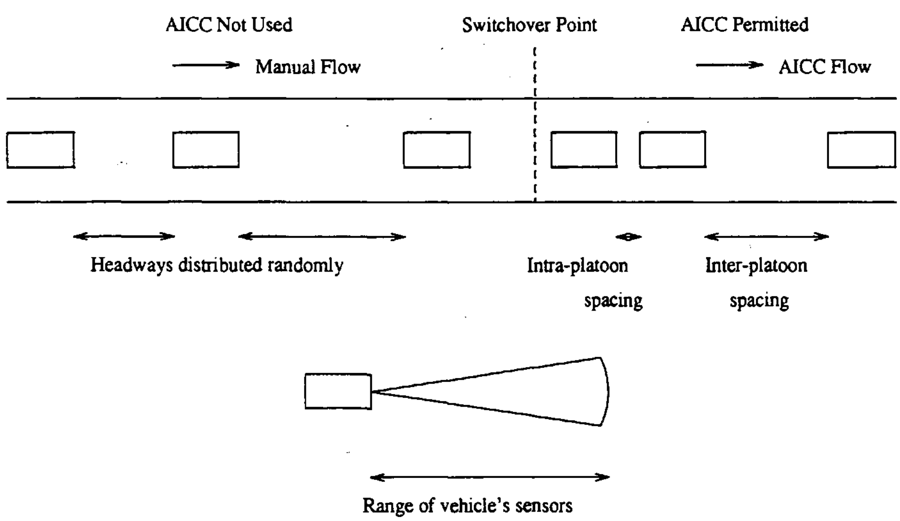


FIGURE 3 How vehicles platoon initially. It is assumed that manually controlled, randomly distributed vehicles enter a section of highway where AICC can be used. Platoons are formed in the AICC section as vehicles detect their predecessors. The lower figure shows the range of detection of a vehicle using on-board sensors.

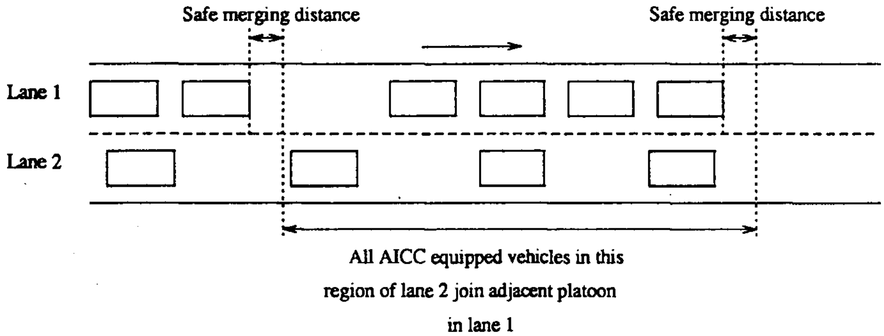


FIGURE 4 The scenario modeled. Vehicles in the transition lane (Lane 2) attempt to enter the AICC lane (Lane 1) and form platoons using AICC. It is assumed that all AICC vehicles within a platoon's catchment region will eventually join that platoon.

Given a flow ϕ_t of cars in the transition lane, regularly spaced from ramp metering, we say that the number of vehicles in the transition lane within the catchment area of a platoon is q:

$$q = \frac{L(N)\phi_t}{3{,}600\upsilon}$$

We say that all AICC-equipped vehicles in a platoon's catchment area eventually will join that platoon so that a new probability distribution for platoon size $p'(N)$ can be derived from the old one $p(N)$:

$$p'(N) = \sum_{n=1}^{N} [\gamma p(n) p(\text{aicc})^{N-n} p(\text{man})^{q-(N-n)}]$$

$$\gamma = \begin{cases} 0 & q - (N - n) < 0 \\ \dfrac{q!}{[q - (N - n)]!(N - n)!} & \text{otherwise} \end{cases}$$

The new distribution is calculated as the probability of ending up with a new platoon of size N starting with a platoon of size n and adding $N - n$ vehicles from an available number q, given a particular p(aicc).

Caveats

After a number of AICC-equipped vehicles have entered the inside lane, the proportion of AICC-equipped vehicles in the inside lane will not be the same as it was in the manual part of the lane; however, since we are attempting to find the maximum flow in a lane we only investigated this case. By allowing vehicles to enter through merging we can assume that any vehicle that has the capability to join a platoon will eventually do so.

We assume several discrete entrance points and that vehicles from these entrances join the inside lane as soon as possible. In reality it may be that drivers enter the highway and that only after some time do they try to enter the inside lane. Relaxation of this assumption would not alter the capacity of the inside lane but may allow a higher flow on the transition lane by spreading out the flow of vehicles wanting to enter the inside lane.

Another drawback of our model is that we do not allow vehicles to enter the inside lane unless it is to join an existing platoon. Formation of new platoons in the inside lane would expedite the process of moving vehicles into that lane.

Results

We investigated a scenario in which a flow of 1,800 vehicles/hr was on the inside lane in the manual section. We set p(aicc) to the proportion of AICC-equipped vehicles initially on the highway. We evaluated the maximum flow that could result from a given distribution of platoon sizes and set $\overline{\Delta} = \overline{\Delta}_{man}$. Each platoon leader behaved as though driving in a manual lane and thus gave rise to the headway distribution found on such a highway lane.

Choice of Δ_{max} can be anywhere between 20 m if the vehicles are going around a bend in the rain at night to 150 m (*2*) if the vehicles are on a straight flat stretch of road in the daytime. Since we are attempting to derive an upper bound, we assume that Δ_{max} is maximized.

There are six entrances, each at a spacing of 1 km with the first at 1 km after the AICC section starts. On-ramp flows are such that AICC vehicle flows from the transition lane to the inside lane constitute a rate of 720 vehicles/hr. We calculated the platoon size distribution after each entrance, assuming that all the vehicles that wanted to enter the inside lane had done so. From these distributions we calculated the maximum possible flow.

Manual flow was calculated by the formula

$$\text{Flow} = \frac{3{,}600\upsilon}{l + \Delta_{man}}$$

and was used as a reference for calculating percentage capacity gains. We take the speed $\upsilon = 25$ m/sec, the length of a car $l = 5$ m, the average headway $\overline{\Delta}_{man} = 45$ m, and the intraplatoon spacing $d = 8$ m (*1*). Despite a previous analysis (*1*), we believe this to be an extremely optimistic value; in reality the intervehicle spacing may have to be higher to allow vehicles to stop without colliding in the case of emergency braking (*5*), but we use this value because we are interested in an upper bound. Because in this model it is not $p(\Delta)$ but $p(\Delta < \Delta_{max})$ that is of importance, we obtain results for a range of values for $p(\Delta < \Delta_{max})$. Note that for the assumptions made, the capacity of the inside lane is at maximum (using Equation 1), (3,600 × 25)/(8 + 5) = 6,900 vehicles/hr. Results are shown in Figure 5.

Obviously a greater number of AICC vehicles at the start allows the flow to build faster through formation of more and larger platoons, but predicted growth of all flows slows down as capacity is approached. This model shows that capacity can be attained but may take a long time to achieve.

DETAILED KINEMATIC SIMULATION TO ANALYZE AICC

The above analyses allowed us to gain some knowledge of likely platoon size distribution and potential flow benefits using AICC; however, because they are fundamentally steady-state analyses they can afford no information on stream stability, likely access rates, or access times for vehicles in such systems. Since merge junctions are likely to be the principal sources of delay and capacity limitation of an (even partially) automated system, it is necessary to progress beyond steady-state analysis and examine transient behavior (*6*).

To study these phenomena we use a modified version of SmartPath (Eskafi) (*7*), a detailed, multiprocess kinematic simulator used to aid design of coordinated platooning strategies. The simulator has been modified to allow accurate representation of the behavior of AICC-controlled and manually driven vehicles (Figure 6). The control policies proposed previously (*1*) have been used to model the dynamics of AICC vehicles.

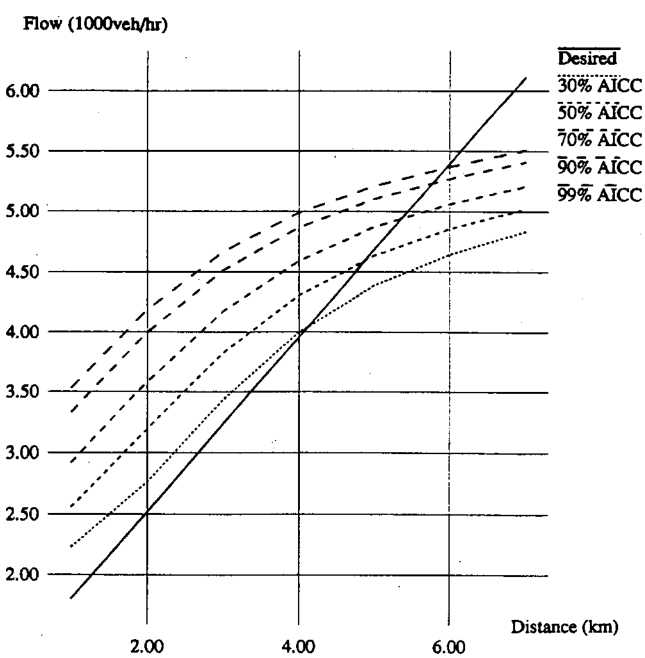


FIGURE 5 Maximum possible flows on the AICC lane given $p(\Delta < \Delta_{max}) = 0.7$ and different initial proportions of AICC vehicles. As the lane approaches theoretical maximum capacity, the obtained maximum flows begin to lag behind demand as it becomes increasingly hard to find the space to fit in more vehicles.

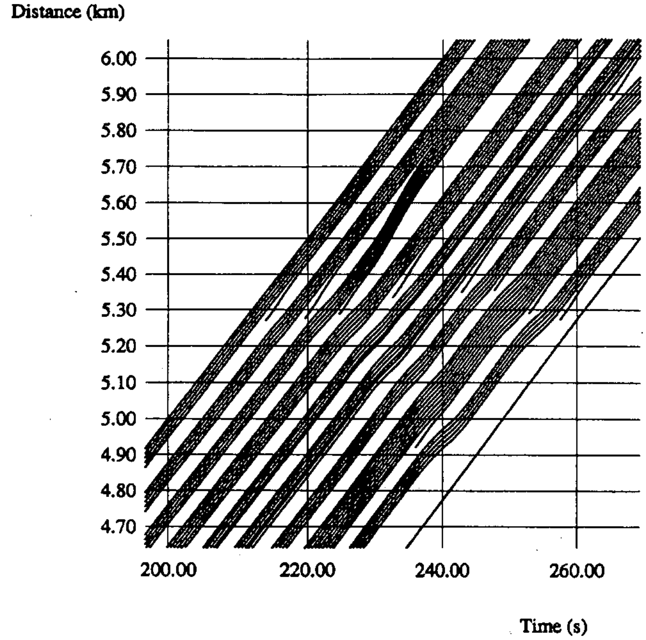


FIGURE 6 Detail of SmartPath output for inside lane showing how vehicles entering the lane join the rear of existing platoons. Opportunistic merging of two platoons is shown.

Scenario

We used the highway configuration shown in Figure 1 and analyzed above. A randomly distributed stream of vehicles at 1,800 vehicles/hr was on the inside lane, traveling at 25 m/sec under manual control. At a distance of 0 km those vehicles that were equipped with AICC used their systems and began to platoon with each other. For this simulation we say that each vehicle can detect up to 60 m ahead. After some distance d_1 the platoons will be in steady state (they will all be separated by at least 60 m). We place the first on-ramp such that all the traffic in the AICC lane is already at steady state by this point. The subsequent on-ramps are also placed sufficiently far along the highway that all the traffic from the previous on-ramp has entered the AICC lane and that the AICC lane is in steady state (Figure 7).

Vehicles that entered from the on ramps were all AICC equipped and immediately tried to enter the inside lane. If there is a sufficiently large (safe) gap adjacent to a car it will enter the inside lane; otherwise it will decelerate until it is far enough behind a platoon to enter the lane and merge with the platoon.

Off-ramps are not included in this simulation because within SmartPath it is not possible to simulate the AICC method of leaving the AICC lane (just pulling out into the transition lane). SmartPath requires the platoon to break and allow a vehicle more room to leave. However, we can still comment on how the AICC exiting strategy may affect stream stability (see conclusions).

Intraplatoon separation was derived in the manner of Ioannou and Chien (*1*) to be 8 m. Interplatoon safety distance was 50 m, and the maximum platoon size was 30. With these parameters the maximum theoretical flow is about 6,250

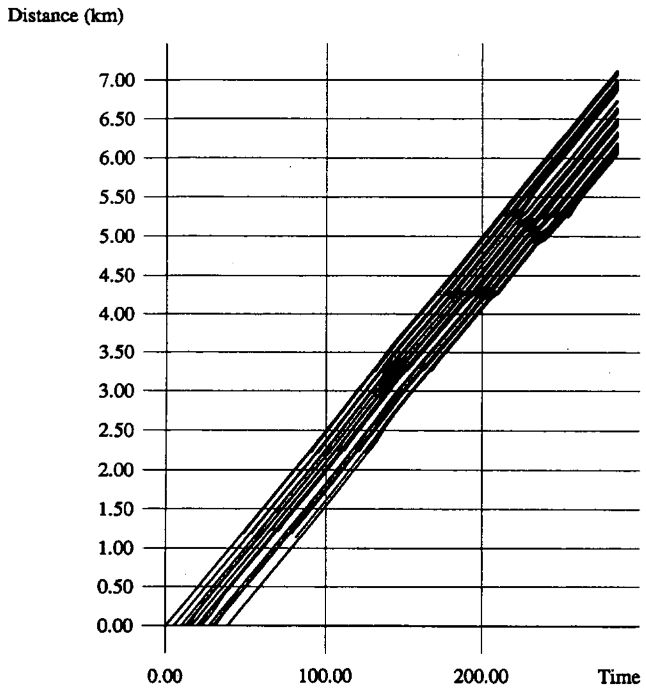


FIGURE 7 Time-distance diagram for the automated lane. Vehicles enter the automated section of the highway at 0 km at a flow of 1,800 vehicles/hr moving at 25 m/sec. There are entrances at 1, 2, 3, 4, and 5 km that release vehicles at the rate of one every 5 sec (720 vehicles/hr). Note how the traffic sample "spreads out" when capacity is exceeded (after 5 km).

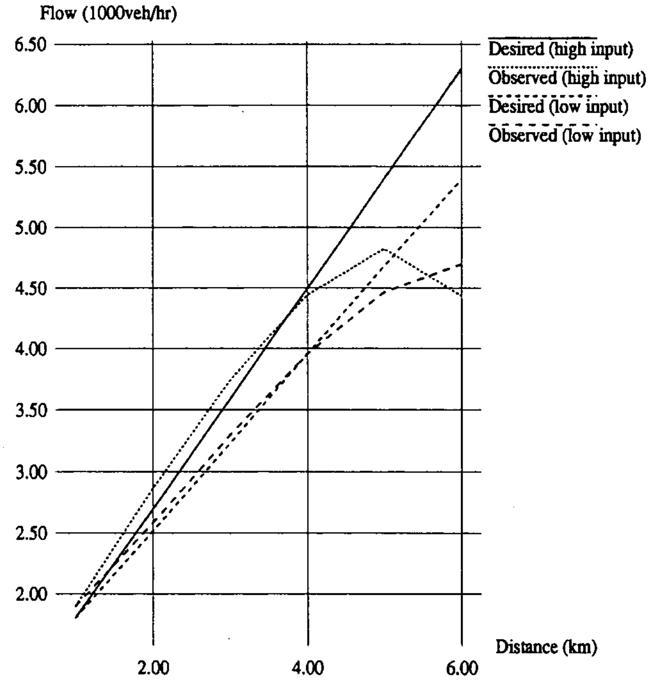


FIGURE 8 After Monte Carlo runs, the average flow on the inside lane at each point along the highway for initial *p*(aicc) = 0.5 is shown. Low on-ramp entrance rates (720 vehicles/hr) track demand better than high rates (900 vehicles/hr).

vehicles/hr. To simulate realistic vehicle dynamics we limit the deceleration of a vehicle during a maneuver to 0.3 *g*.

For different on-ramp flows and at each stage of the highway we attempted to measure the following:

- The maximum attainable flow in the inside lane,
- The maximum sustainable on-ramp flow rate, and
- The effect on stream stability of vehicles entering.

Because all of these are random variables dependent on the initial headways of manual traffic in the inside lane and on exactly how platoons are formed, we obtained our results after Monte Carlo simulations to attempt to average out the effects of the random behavior. The simulation was run 100 times with different starting headways (all drawn from the same distribution), and the cumulative results were analyzed.

Experimentation

To allow results to be obtained within a reasonable time frame we generated a 40-sec section of traffic and monitored the progress of this "sample" as it proceeded along the highway. In the first instance we wished to determine roughly the rate at which vehicles can be input from an on-ramp. Initial experiments showed that when the flow on the AICC lane was high the vehicles in the transition lane had to maneuver significantly to change lanes, resulting in turbulent flow in the transition lane and some vehicles not being able to change lanes at all. It was found that on-ramp flows of 720 vehicles/hr produced the best results, maximizing eventual maximum flow in the inside lane without excessive turbulence (Figure 8). Monte Carlo simulations were therefore conducted using on-ramp flow rates of 720 vehicles/hr.

Results

Figure 9 shows the actual flow rates achieved in the AICC lane. These results are based on averages over the entire Monte Carlo sequence and show that the observed flow rate is lower than might be expected from our model. Initially, when traffic density is low, platooning causes the traffic on the lane to bunch up and thereby increases the measured flow rate over the traffic sample. However, as the density on the AICC lane increases it is found that the length of the traffic sample grows to accommodate new vehicles, thereby effectively reducing the increase in flow. Growth in the flow rate therefore slows down as flow on the AICC lane increases as expected.

Obtaining quantifiable results on stream stability is not easy. From the experience of running these simulations we can make the following broad comments. With regard to the AICC lane, it appears that the flow is stable under conditions of vehicles entering the lane, until lane capacity is approached. Vehicles leaving platoons also may not cause perturbations as long as the sudden change in headway for a vehicle following a vehicle about to exit does not cause it to behave erratically. The real source of flow instability in AICC systems is found in the transition lane. Note also that in these simulations there was no permanent flow on the transition lane—

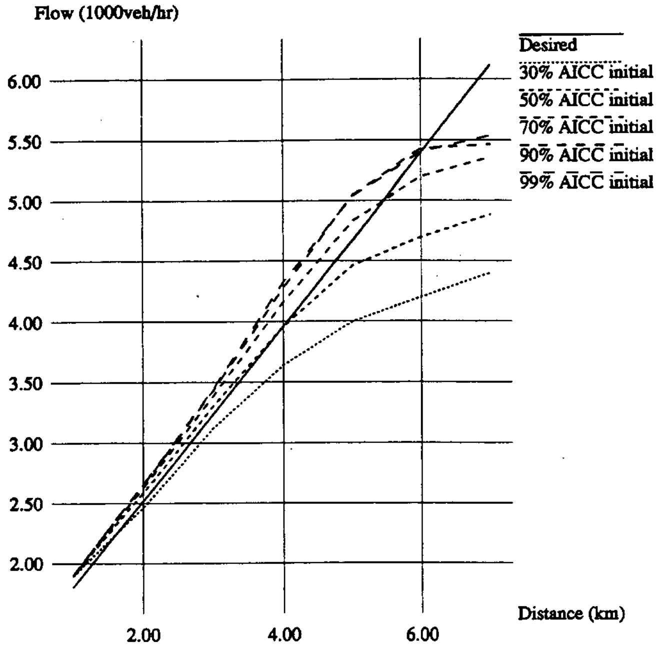


FIGURE 9 Observed flow at regular intervals along the highway using on-ramp flows of 720 vehicles/hr and given various proportions of AICC-equipped vehicles in the initial stage. Demand flow is tracked well until the theoretical capacity of the lane is approached.

only vehicles attempting to join the AICC lanes. Introduction of even a small permanent flow would have severe effects on stream stability, and the maximum on-ramp flow would be curtailed.

Faults in the Simulation Methodology

The above results were obtained using a modified version of SmartPath. This allowed us to investigate AICC quickly, but because the program is designed for another IVHS strategy it implies that not all of the features of AICC can be modeled accurately. The following is a list of areas in which this simulation does not do justice to the concept of AICC:

- We assume that a driver in the transition lane who is adjacent to a platoon and wants to enter the inside lane will take positive action and decelerate to join the back of the platoon. In reality it may be that a speed differential (albeit small) may exist between the two lanes, allowing drivers in the transition lane to simply wait until a safe gap rolls past them and allows them to change lanes. This behavior would reduce the turbulence in the transition lane but may increase the amount of time it takes drivers to enter the inside lane. If, however, the use of AICC is perceived as providing a high degree of comfort, drivers may try to enter a lane where its use is permitted as soon as possible.
- In AICC a vehicle should simply enter the AICC lane if there is a gap; if there is not a gap, a vehicle should slow down until one appears. In SmartPath this logic has been implemented, but a vehicle that is slowing down to join the rear of a platoon is also in communication with it. Because a platoon is able to communicate to only one vehicle at a time, no other vehicle is able to talk to that platoon, which may create cases in which vehicles do not change lanes in a realistic order. However, in the majority of cases this will not occur because the vehicles forced to slow down behind a decelerating vehicle in the transition lane will come across a suitable gap first and merge sooner into the AICC lane on their own.
- One area in which SmartPath is currently deficient is the control law for the lead vehicle of a platoon. (To our knowledge no clear control law for lead vehicles in AICC platoons has been proposed. Merely maintaining a constant speed may not be sufficient to cope with all eventualities.) This law calculates the deceleration a vehicle must apply at each instant simply on the basis of the distance and relative velocity of the vehicle in front so that a collision is always avoided (*8*). The same law is also used for vehicles in the transition lane, and its crude nature leads to situations of traffic stream instability that would not occur with manual drivers. This may have led to a reduction in the maximum on-ramp flow, but at high flows on the AICC lane the problems of large transition delays will remain no matter what control laws are used.
- SmartPath cannot simulate the AICC policy for exiting, which is simply to pull out. However, an exiting policy such as this may not have adverse effects on stream stability in the AICC lane but may cause disruption in the transition lane, especially if there is some permanent flow in that lane.

These drawbacks notwithstanding, some strong conclusions about AICC can still be made from use of the simulator.

Conclusions

AICC seems to provide stable flow and an increased capacity in the automated lane (albeit on the basis of an 8-m value for *d*, which may not be realistic), but it has several drawbacks:

- Building up the flow on the AICC lane takes a long time and perhaps much longer than with coordinated platooning, which can sustain higher on-ramp flows even when flows on the automated lane are very large (see companion paper by Rao et al. in this Record).
- Flow on the transition lane is not very stable when the flow on the AICC lane becomes large.
- The theoretical upper limit on the capacity cannot be attained in any reasonable length of time, and a practical limit on the attainable flow appears to be about 75 percent of the theoretical flow.
- Because the stream in the transition lane is so sensitive to vehicles moving in and out of it, it may prove necessary to ensure that only a very small or even no permanent flow exists on that lane. This implies the need for two lanes with only one in which traffic is actually traveling permanently. Therefore any capacity on the AICC lane must be divided by 2, and if the practical limit is 5,500 vehicles/hr the overall highway flow is 2,700 vehicles/hr per lane, which is only marginally better than highways maintain now. This point is crucially dependent on our model of human vehicle-following behavior, which is admittedly deficient in this simulation. However, given high flow rates it seems reasonable to assume that disturbances in the transition lane are inevitable regardless of the vehicle-following models used.

CONCLUSIONS

Proponents of AICC technology claim significant increases in capacity, but the claim has not been substantiated by convincing analysis in the published literature. We have attempted to provide such an analysis using theoretical models that ignore interaction among vehicles as they attempt to merge and the SmartPath simulation system that does model those interactions.

It is found that the estimates for maximum capacity increases using both approaches agree when the level of penetration of AICC technology is under 40 percent; but at higher levels of penetration the theoretical model overestimates the capacity increase.

The simulator, SmartPath, was then used to study the more difficult problem of stream stability, especially in the transition lane used by AICC-equipped vehicles entering from on-ramps and intending to use the AICC lane. It is found that as the flow entering from the on-ramps begins to reach the values estimated in the previous analysis, it becomes increasingly difficult for vehicles to enter the AICC lane. Taking published values of AICC control strategy, we find that the AICC lane can achieve a flow of 5,500 vehicles/hr—three times the current capacity. However, if this flow is to be achieved, it is not possible simultaneously to have any permanent flow in the transition lane, so that the effective flow per lane is reduced to 2,700 vehicles/hr.

Although these flow rates appear impressive, they are based on an extremely optimistic value for intervehicle spacing in platoons using AICC. This value (8 m) is, in our opinion, too small and is unlikely ever to be implemented for the following reason: If a vehicle is equipped with AICC but has poor brakes or tires, it will be a hazard to any vehicle it follows that is in better condition mechanically (*5*). If a sudden deceleration is required, the vehicle in relatively poor condition will not be able to stop before hitting its predecessor. Apart from the safety issue, this raises the important legislative question of liability.

Because AICC cannot maintain the ultraclose spacings (1 m) required to prevent high relative velocity collisions [as defined previously (*9*)], and the proposed spacings of about 8 m could lead to high relative velocity collisions, we are forced to conclude that spacings comparable with those used by human drivers now will be required for safe operation. This immediately reduces the potential for capacity gain from use of AICC to very low levels, although the potential for increase in driver comfort still exists.

The argument presented here is tentative, and we have made explicit several assumptions that underlie both the theoretical and simulation approaches. We hope the argument will stimulate more careful assessments of the AICC technology.

ACKNOWLEDGMENTS

This work is supported by the California PATH program and California Department of Transportation. The authors would like to thank F. Eskafi, S. Sheikoleslam, S. Shladover, and R. Hall for many useful comments and helpful guidelines.

REFERENCES

1. P. Ioannou and C. C. Chien. *Autonomous Intelligent Cruise Control*. Technical Report 92-05-01. Department of Electrical Engineering—Systems, University of Southern California, May 1992.
2. R. Muller and G. Nocker. Intelligent Cruise Control with Fuzzy Logic. In *IEEE Conference on Intelligent Vehicles*, Detroit, 1992, pp. 173–178.
3. F. Broqua, G. Lerner, V. Mauro, and E. Morello. Cooperative Driving: Basic Concepts and a First Assessment of Intelligent Cruise Control Strategies. *First Drive Proceedings*, 1991, pp. 144–147.
4. A. D. May. *Traffic Flow Fundamentals*. Prentice Hall, 1990.
5. A. Hitchcock. Coefficients of Friction and Automated Freeways. PATH MOU 19 Occasional Paper 1. May 1992.
6. S. Shladover. Operation of Merge Junctions in a Dynamically Entrained Automated Guideway Transit System. *Transportation Research*, Vol. 14A, 1979, pp. 85–112.
7. F. Eskafi and P. Varaiya. *SmartPath: An Automated Highway System Simulator*. Technical Report 92-0708-1. California PATH, 1992.
8. F. Eskafi, D. Khorramabadi, and P. Varaiya. *SmartPath: Automatic Highway System Simulator*. Preprint. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, June 1992.
9. P. Varaiya and S. Shladover. *Sketch of an IVHS Architecture*. Technical Report UCB-ITS-PRR-91-3. Institute of Transportation Studies, University of California, Berkeley, 1991.

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Automatic Speed Monitor: An Intelligent Vehicle Highway System Safe-Speed System for Advance Warning or Hazardous Speed Monitoring

Alan R. Kaub and Thomas Rawls

In the development of a safer highway driving environment, unsafe speeds have been found to be a cause of approximately one-third or more of all accidents. Monitoring of safe speeds has been commonly performed by radar enforcement. However, recent experience has indicated that long-term exposure to the radar emissions may pose a hazard for the law enforcement community. A new, simple and inexpensive intelligent vehicle highway system speed monitoring concept that can be easily adapted to existing vehicles and roadways is presented. The concept relies on the speed-distance-time relationship and an on-board impulse detector and constant timer to calculate the travel time or posted speed of the roadway. The calculated travel time or speed can then be compared with the current vehicle speed or with a constant time to warn the driver of hazardous approaching events (excessive speeds approaching construction or maintenance work sites, sharp curves, railroad crossings, or school zones). It can also be used to issue tickets to the vehicle registration if warnings remain unheeded. Current estimates indicate that the use of the speed monitoring device may save approximately 10,000 lives and 500,000 injury accidents per year. The cost of the on-board electronics for a magnetic detection system is approximately $10 per vehicle with roadway detection costs for magnets of under $10 per installation. An on-board magnetic sensing device has been successfully tested in field trials to warn of approaching sharp curves and awaits further development and testing of other applications and human factor considerations.

Each year in the United States alone, almost 40,000 persons die in automobile-related crashes, with more than 1,700,000 receiving disabling injuries. Even with new advances in the use of air bags, nonlocking brakes, mobile and air emergency medical service, and similar techniques, the total volume of fatalities and injuries remains constant because of the continuing increase in exposure by the ever-increasing amount of travel performed on the highway system. Research indicates that inappropriate speed is cited as a cause in over one-third of all fatal accidents and by implication in a similar number of injuries and total accidents (*1*). If driver speed can be maintained at safe levels either by providing warning assistance to the driver or by using stricter enforcement techniques applied to the vehicle registration itself, it may be possible to reduce total accidents by one-third (those caused by excessive speeds), which may save more than 10,000 lives and 500,000 disabling injuries each year.

Civil Engineering and Mechanics Department, University of South Florida, 4202 Fowler Boulevard, Tampa, Fla. 33620-5350.

Technology for providing speed warning or enforcement has changed little since the 1930s and 1940s. In that era as today, the major technique to control speeds rested with the provision of either regulatory or warning signs to the driver, who compared the on-board speedometer with the roadside signing and in theory acted appropriately to conform to the signing. Where the driving speeds remained inappropriately high, enforcement by local authorities through the use of "car following" or radar speed monitoring and the issuance of tickets was and still is the primary resource to control high operating or erratic speeds. These techniques remain the primary methods of enforcement today, although video speed monitoring of passenger vehicles and even satellite speed monitoring of common carriers are helping maintain a safer driving environment. However, the costs of such monitoring in economic terms and even lost personal freedoms (video surveillance) are issues that remain to be addressed.

INTELLIGENT VEHICLE HIGHWAY SYSTEM AUTOMATIC SAFE-SPEED CONCEPT

The acquisition and use of information by a vehicle to assist in driver decision making requires either the presentation of information from a stationary roadway site to a moving vehicle or acquisition of information by the vehicle as it passes a specific site. An example is the application of a blanket radio signal that all vehicles receive or the vehicle itself searching out a particular frequency for information. For the purposes of accuracy and reliability, the concept of in-vehicle acquisition is clearly a more timely and reliable vehicle data acquisition system because it eliminates extraneous signals and searches only for a particular impulse. One of the most likely types of systems is an electronic system that transmits continuously from the roadside to the vehicle with limited distance, directional antennae where impulses may be given to independent vehicles in independent lanes from either overhead or roadside locations. A second, less-expensive impulse that the vehicle may receive independently is a magnetic impulse from each lane of the pavement itself. The research contains little information on the use of permanent magnets in intelligent vehicle highway system (IVHS) applications, probably because placing permanent magnets in pavement lanes may bear tort liability implications if the permanent magnets are sufficiently strong to retain nails or magnetic debris in the lane. However, even with this limitation caused by tort liability concerns, if the magnetic field can be maintained at sufficiently weak levels, the magnetic transmission concept may be as sound as the use of electronic transmission media.

When permanent magnets are embedded transversely in a pavement structure, a vehicle can pass over a magnetic field at any speed and still receive the most timely information that the vehicle is capable of receiving. By using two similar permanent magnets embedded a specific distance apart in the pavement (or from overhead in a particular lane), the vehicle can receive two independent impulses. By comparing the time displacement between the two roadway impulses with an onboard constant timer, the vehicle has the ability to calculate the proper speed of the roadway from the following standard time–distance relationship:

$$\text{Distance} = 1.47 * V * T$$

where

D = distance between roadway magnets sets (ft),
V = posted speed of the roadway (mph),
T = constant time (sec), and
1.47 = approximately 5,280 ft/3,600 sec/hr.

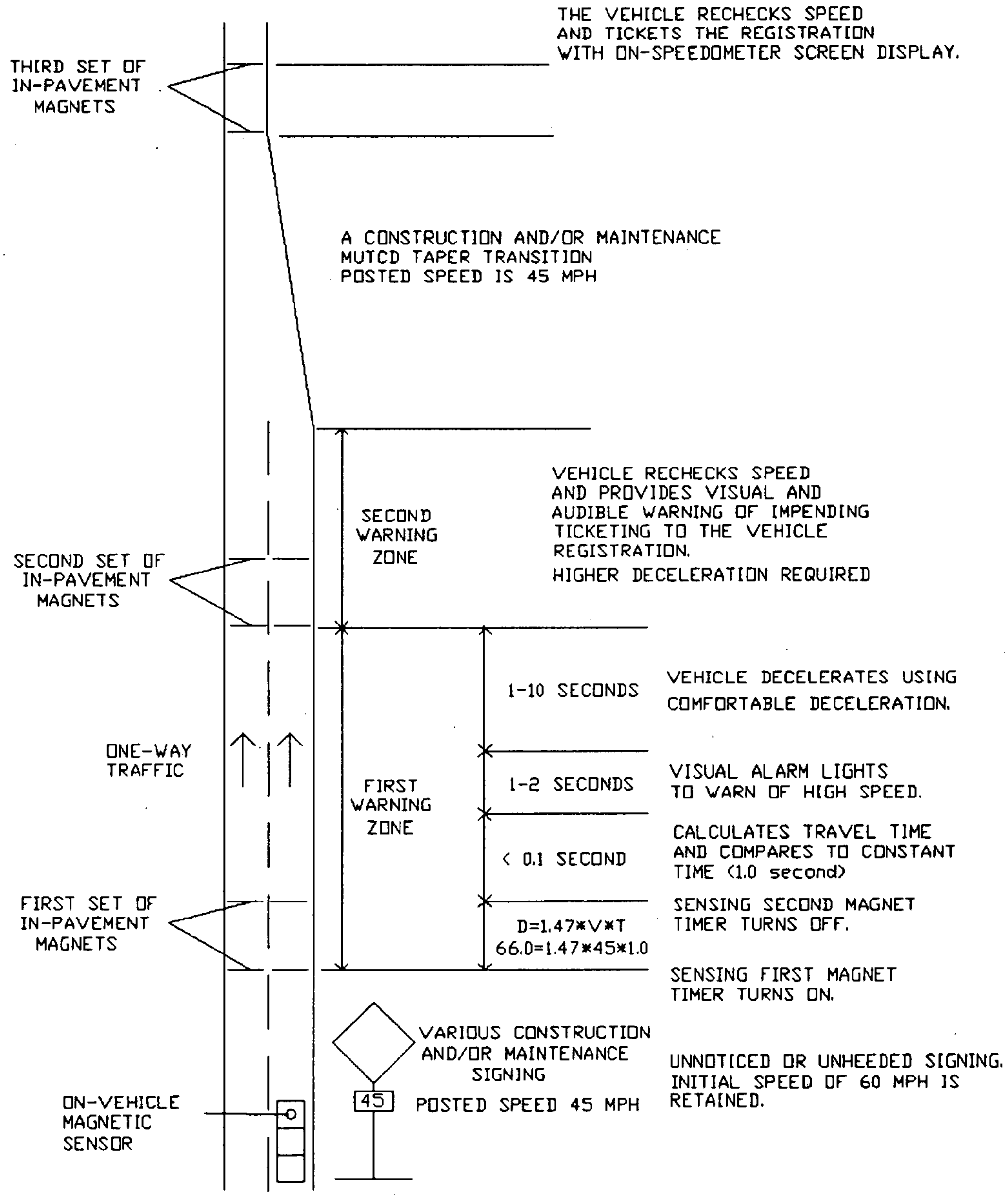


FIGURE 1 General concept of IVHS automatic safe speed system.

Using this relationship, the vehicle itself can calculate the posted speed of the roadway by using the time displacement of the two magnetic impulses. In a similar manner, where a constant timer (imputing a constant time—*T*) is used and the distance on the ground is varied to conform to posted speed limits, the vehicle can also determine whether the actual time to traverse the distance between the magnets is too short (speed too high) or the time between the impulses is below a specified threshold, indicating actual speed below the posted speed limit. Used in this manner, the vehicle is not testing speed but the surrogate of time displacement, which is compared with a time preset into a microchip on board the vehicle. As shown in Figure 1 (assuming a preset microchip timing device of 1 sec on board the vehicle), a vehicle is approaching a construction or maintenance work zone, which is speed restricted to 45 mph; if the vehicle is to traverse the site at or below 45 mph, the magnets must be placed 66.1 (1.47*45*1) ft apart. If the vehicle traverses this distance in less than 1 sec, the speed must be above 45 mph, and if the time to traverse the 66.1 ft is greater than 1 sec, the vehicle must have traversed the segment at less than 45 mph. In this manner, the vehicle has the ability to ascertain for itself whether the driver is acting properly in advance of a particular site where the magnets are placed in the pavement. Sensitivity of the warning and ticketing function (and the lack of precision in sensing the magnetic field) can also be addressed by using a distance greater than 66.1 ft to represent a factor of safety.

The operation of this concept may be triggered by a variety of devices that may be implanted in the roadway, including standard electromagnetic loops powered by on-site electricity, or may even include optical spectrometers to sense vehicular presence. However the information is provided to the vehicle, an on-board electronic microtimer receives the information and, using the above speed-distance model, can determine through a variety of comparative means whether the vehicle is above the desired speed in advance of a hazard and then relay that information to the driver through visual or audible or a combination of visual and audible signals to alert the driver of inappropriate driving behavior with respect to speed. Ultimately, an in-vehicle sensor can be given the capability to issue warning or actual speeding tickets to the owner of the vehicle through the use of speed or odometer "cover-up" devices (much like an "emissions" flag over the odometer) if the vehicle continues to violate the maximum speed. In this instance the speeding tickets recorded in the vehicle memory may be issued when the next regular emissions check is made or at the vehicle registration when the vehicle ownership is transferred.

Hazardous speed locations that may benefit from this application may include both rural and urban driving, but because of the speed of rural operations, it may be expected that such an advance-speed-warning IVHS system would offer greater benefit where the speed and therefore severity of the accident can be reduced. In the rural environment this may occur at isolated stop locations on high-speed roadways, on sharp, high-speed curves, locations of poor skid resistance, railroad crossings, or a variety of other locations involving high-speed roadway or roadside accidents. In the urban environment, this need for speed reduction may occur at a variety of similar locations such as hazardous intersections, neighborhood speed restraints, school zones, or even sight restrictions on vertical crest freeway curves subject to peak-hour queuing. However, probably one of the most productive applications will be at construction and maintenance work sites where speed monitoring and control are almost impossible to perform and yet where approximately 80 percent of accidents occur (*2*).

At present, an on-board magnetic detector and timers have been produced at a cost of under $10, with the permanent pavement magnets also purchased for under $10. The magnetic concept using permanent magnets laid on the pavement surface was field tested at several sharp curve sites in Tampa, Florida, and performed as expected by warning the driver of excessive speeds approaching sharp curves.

Future efforts, including human factors testing of visual or audible warnings systems and automatic vehicle speed ticketing mechanisms await the interest of FHWA, NHTSA, or the IVHS industry. Most important, if inappropriate speed is responsible for an estimated one-third of all fatal accidents as reported, it may be expected that federal insistence on the introduction of the automatic speed monitoring device to the vehicle has the clear potential to save over 10,000 lives and over 500,000 disabling injury accidents, and, by assumption, a similar amount of property damage accidents each year, while also freeing the law enforcement community from the potential hazards of radar speed enforcement.

REFERENCES

1. *Accident Facts—1989*. National Safety Council, 1990.
2. B. T. Hargroves and M. R. Martin. *Vehicle Accidents in Highway Work Zones*. FHWA/RD-80-063, 1980.

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Engineering Feasibility of Roadway Electrification in a High-Occupancy-Vehicle Facility

T. CHIRA-CHAVALA AND EDWARD H. LECHNER

The preliminary engineering feasibility for early deployment of a roadway-powered electric vehicle in El Monte Busway (a 3+ high-occupancy-vehicle facility in Los Angeles) is assessed. The evaluation consists of the determinations of the scale of electrification, locations to be electrified, mode of operation, level of energy transfer, and energy consumption. These analyses are accomplished through vehicle simulation. The evaluation results indicate that systems that combine the electrified roadway and static chargers or those that use static chargers exclusively can be deployed in El Monte Busway. A plan for the technology demonstration is developed. Hardware costs for this plan are estimated.

Concern over air quality in California has led the state to enact a law requiring that 2 percent of all vehicles sold by 1998 and 10 percent sold by 2003 be "emission-free" vehicles. Electric vehicles (EVs) are generally considered to be emission free because they do not emit pollutants while running on the road or stopping in traffic, although power plants supplying electric power to them do emit pollutants. Large-scale use of EVs has not materialized because they have limited range between battery recharges; in addition, the existing battery technology requires 6 to 8 hr for a full recharge. One way to increase the range of Evs between overnight battery recharging in the absence of advanced battery technologies is through the use of roadway-powered electric vehicles (RPEVs).

RPEVs are hybrid electric-electric vehicles (*1–3*) that use an inductive coupling power transfer principle, in which energy stored in an on-board battery is supplemented by energy transferred to the vehicle through an inductive coupling system (ICS). The ICS uses a magnetic field to transfer power across an air gap from the roadway inductor (buried underneath the pavement) to the vehicle. This system can be thought of as a large transformer with an air gap. The primary source of this transformer is the roadway inductor, and the secondary source is the pickup inductor. RPEVs can operate both on and off electrified roadways. On the electrified roadway, they draw power directly from the roadway for use in propelling the vehicle. The balance of this roadway power that is not used is stored as a reserve in the on-board battery. Off the electrified roadway, the vehicles rely solely on the on-board battery power. Electrified roadways do not present electric shock hazards and thus can be shared by pedestrians and nonelectric vehicles. The RPEV technology can be applied to both transit buses and private passenger vehicles.

As part of RPEV research at California PATH and the Playa Vista Project in southern California, a prototype bus and a G-van (a full-sized van) were built and tested. The prototype bus is a low-frequency system with high roadway current (400 Hz and 1,200 amp-turns, respectively). The prototype G-van is a high-frequency system with low roadway current (8,500 Hz and 240 amp-turns, respectively). Design engineers now believe that high-frequency systems with low roadway current are more suitable for the highway application. To advance this technology toward full maturity, a case study was initiated to determine the engineering feasibility of early RPEV deployment in El Monte Busway, an existing 3+ high-occupancy-vehicle (HOV) facility in Los Angeles.

STUDY OBJECTIVE

The objective of this paper is to evaluate the preliminary engineering feasibility of early deployment of the RPEV technology in El Monte Busway and to suggest a plan for early technology demonstration in this HOV lane.

DESCRIPTION OF EL MONTE BUSWAY

El Monte Busway is an exclusive-access HOV lane, not an exclusive bus lane as its name may imply. It is located on I-10 (the western end of San Bernardino Freeway). It is about 18.6 km long, with its western end originating in downtown Los Angeles (at Alameda Street). To the est, the HOV lane ends at the El Monte bus terminal. This HOV lane has one travel lane in each direction, with the two directions separated from each other by permanent barriers. For about 4 mi at the western end, the HOV lane is separated from the freeway mainline by permanent barriers. For the remaining length, the HOV lane is separated from the freeway mainline by buffers at least 1.2 m wide. Access and egress to El Monte Busway are through exclusive ramps.

Being a 24 hr HOV facility, El Monte Busway opens to private passenger vehicles and buses with at least three occupants. Peak traffic flow in the busway is over 1,200 vehicles per hour per lane. There are over 500 bus trips per day in both directions combined.

T. Chira-Chavala, Institute of Transportation Studies, University of California, 109 McLaughlin Hall, Berkeley, Calif. 94720. E. H. Lechner, Systems Control Technology, Inc., 2300 Geng Road, Palo Alto, Calif. 94303.

METHODOLOGY

Although roadway electrification can be applied to private passenger vehicles, transit buses are more likely to be early users of this technology, particularly before the critical mass of the electrified roadway is achieved. Therefore, transit buses are selected as the design vehicle in this paper. The engineering feasibility evaluation of RPEV in El Monte Busway involves the determinations of functional requirements for a proposed system, electrification scale and energy transfer level, energy consumption, and preliminary design of the ICS. The analyses of the electrification scale, energy transfer level, and energy consumption are accomplished through vehicle simulation, using an electric-vehicle simulation model called EVSIM (*4*). The vehicle simulation requires the following data input:

1. The following information on bus lines currently operating on El Monte Busway was compiled: route length and layout, scheduled bus travel time and stops, dwell time at bus stops, layover time, travel speed, and roadway grades. Nine bus lines were analyzed (Table 1). The posted bus numbers were recoded in ascending order of route length (using numbers 1 through 9), as shown in the last column of Table 1.
2. Characteristics of the hypothetical El Monte bus are as follows: gross vehicle weight, 13 930 kg; maximum acceleration, 3 m/sec^2; drag coefficient, 0.60; rated motor base speed, 1,600 rpm; gear ratio, 28 rpm/km/hr; battery voltage, 500; and battery amp-hour rating, 400.
3. The roadway excitation of the El Monte system is assumed to have a frequency of 8,500 Hz.

FUNCTIONAL REQUIREMENTS FOR RPEV SYSTEM

An RPEV system for early deployment in El Monte Busway should have the following functional requirements:

1. It should provide adequate energy for transit buses to operate continuously for at least 15 hr/day between overnight battery rechargings.
2. The on-board battery should not be discharged below 80 percent at any time during normal daily operation. The purpose of this depth-of-discharge (DOD) limit is to maximize the battery life as well as to provide reserve capacity for the buses to operate on the battery power alone for a couple of hours in case of power failure.
3. The ICS output current required should be limited to 300 amp (at 500 V). This limit, based on knowledge gained from RPEV research at California Partners for Advanced Transit and Highways (PATH) and Playa Vista, aims to minimize roadway loss, system costs, and environment concerns over the electromagnetic fields (EMFs).

ANALYSIS RESULTS

Results are presented for the roadway electrification scale; roadway locations to be electrified, mode of operation (dynamic or static charging), ICS output current, and energy consumption. The dynamic mode of operation refers to power transfer from the electrified roadway to the vehicle that occurs while the vehicle is moving. The static mode of operation uses relatively short sections of static chargers (3 to 5 m long), and the energy transfer occurs only while the vehicle is stationary atop these static chargers. Static chargers can be installed at bus stops and layover points to provide quick battery recharges during routine stops and layovers. Static chargers can be an important energy source for roadway-powered electric buses, because 1 min of static charging could provide energy equivalent to that drawn from 1 mi of the powered roadway. Short discrete sections of static chargers are much less expensive than roadway electrification.

Scale of Electrification for El Monte Busway

To determine the amount of roadway electrification required, three electrification scenarios (A through C) are evaluated, in which the travel lanes of El Monte Busway are electrified and static chargers are installed at selected locations. These three scenarios examine operation of existing bus lines that currently use El Monte Busway with no changes to routes or

TABLE 1 Characteristics of Existing Bus Lines Using El Monte Busway

Actual Route Number	Destination	Round Trip Length (kms)	Round Trip Time (including layovers) (minutes)	Total Layovers (minutes)	Average Velocity (including layovers) (km/hr)	Average Velocity (excluding layovers) (km/hr)	Recoded Bus Number*
480	Pomona Express	109	192	39	34.1	43.0	5
491	Sierra Madre	70	123	14	33.9	38.2	1
481	West Covina	83	157	25	31.8	37.8	2
497	Montclair	120	212	34	34.1	40.5	6
486	Puente Hills	86	180	23	28.6	32.6	3
482	Pomona	141	297	54	28.5	34.9	8
484	Ontario Airp.	139	318	60	26.2	32.3	7
488	Glendora	108	259	39	25.1	29.6	4
490	Fullerton	153	296	40	31.0	35.8	9

* Recoded bus number is in an ascending order of the route length

schedules; new roadway-powered electric buses would simply replace the existing diesel buses. This is viewed as the preferred implementation. Two more scenarios (D and E) in which static chargers are to be used exclusively, without electrifying the travel lanes of El Monte Busway, are also evaluated. The first three scenarios are evaluated first and described below.

Scenario A: Electrify El Monte Busway and Install Static Chargers at Bus Stops and Layover Points

Scenario A proposes an extensive electrification scale. It involves electrifying the entire length of the travel lanes in both directions of El Monte Busway. In addition, static chargers will also be installed at bus layover points of each bus line (i.e., the origin and destination of the bus line), as well as at downtown bus stops used by these bus lines.

Scenario B: Electrify El Monte Busway and Install Static Chargers at Bus Layover Points

Scenario B proposes a less extensive electrification than Scenario A. It involves electrifying the entire length of the travel lanes in both directions of El Monte Busway as well as installing static chargers at the layover points of each bus line. However, static chargers will not be installed at downtown bus stops used by El Monte buses.

Scenario C: Electrify El Monte Busway Without Using Static Chargers

The electrification scale proposed for Scenario C is less extensive than those for Scenarios A and B. Scenario C involves electrifying the travel lanes in both directions of El Monte Busway without installing static chargers at the bus layover points or downtown bus stops.

To be considered viable for implementation, each scenario must be capable of providing adequate energy for all bus lines to operate continuously for at least 15 hr/day, without the battery DOD exceeding 80 percent at any time during normal daily operation. Further, the ICS output current required should not exceed 300 amp. The simulation results from the EVSIM model indicate that Scenario A could meet this daily range requirement for all bus lines currently operating on El Monte Busway. On the other hand, Scenarios B and C are found to fall short of meeting this requirement because both scenarios require ICS output current greater than 300 amp to meet the 15-hr daily range requirement for a number of longer bus lines. Therefore, Scenarios B and C are excluded from further discussion in this paper. Detailed results for Scenario A are presented below.

Battery DOD Profile for One Round-Trip Under Scenario A

Bus Line 3 is used to illustrate the simulation results on the battery DOD profile obtained from the EVSIM model. The battery DOD profile for Line 3 for one round-trip under Scenario A is shown in Figure 1. The vertical axis represents battery discharge, and thus a higher positive value indicates that the battery actually has less charge available. Two battery discharge profiles are shown—one for the ICS output current of 250 amp and the other for 300 amp. The figure indicates that, other things being equal, the magnitude of the battery DOD will decrease as the ICS output current increases.

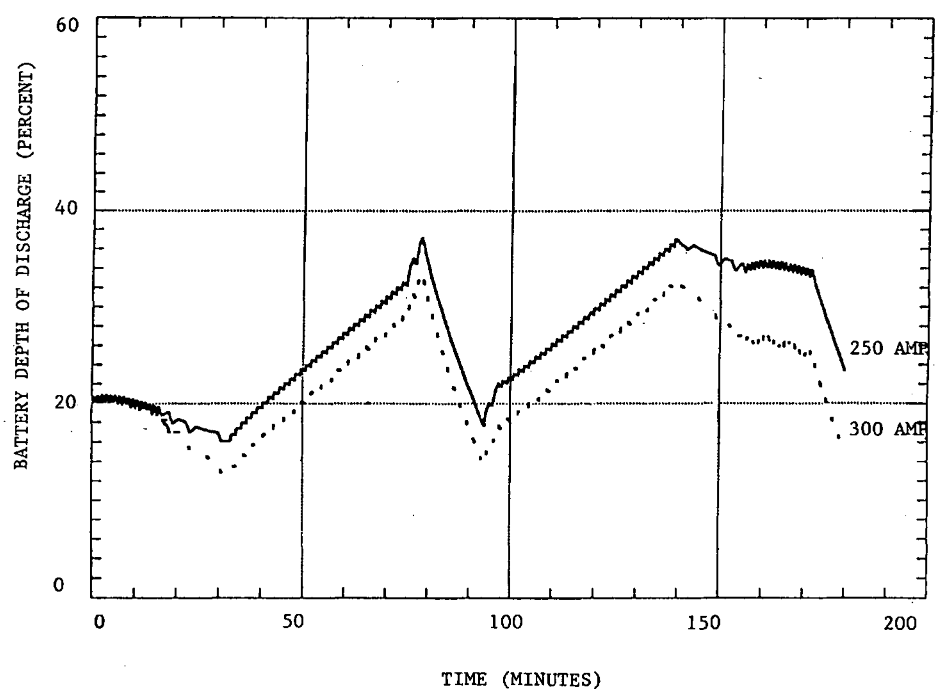


FIGURE 1 Battery discharge profile for Bus Line 3.

Figure 1 shows a battery DOD profile for any one round-trip for Bus Line 3. This particular round-trip starts in the downtown area with 20 percent DOD. The bus traverses downtown streets and then goes onto El Monte Busway (Minute 0 to Minute 30). Figure 1 indicates that there is slight net charging of the battery both in the downtown area (because of static chargers) and on El Monte Busway (because of the electrified lane), indicating that more energy is transferred to the bus than is used. Between Minute 31 and Minute 77, the bus traverses the unpowered suburban portion of the route. At the end of the line, there is net discharge of the battery, which reaches a peak DOD of 37 percent (or 17 percent higher than the DOD at the start of this particular round-trip). Figure 1 also shows that there are very short intervals of battery charging (decreases in the battery DOD) every block because of regenerative braking. This is evident by the ripples, which are farther apart in the suburb than in the downtown, indicating longer distances between stops in the former. The bus then stays idle at the layover point from Minute 78 to Minute 93 while the battery receives a recharge from the static charger. Next, the bus starts the trip back toward the downtown by traversing the suburban portion of the route between Minute 94 and Minute 140. During this time, there is battery discharge until the bus reaches El Monte Busway. From then on, there is net battery charging on El Monte Busway and downtown streets. Finally, at the downtown layover point (Minutes 172 to 180), there is again net battery charging from the static charger at the layover point.

Battery DOD profiles for the other eight bus lines all exhibit patterns of battery charging and discharging for the downtown area, El Monte Busway, the unpowered suburban portion of the route, and layover points similar to that for Bus Line 3, with the actual DOD values varying from bus line to bus line.

Power Transfer Level Required for Scenario A

Consider Figure 1 with the output current of 250 amp. The net change in battery DOD for Bus Line 3 in one round-trip (defined as the difference in DOD levels between the end and the start of that round-trip) is about 3 percent. Therefore, at the start of the next round-trip, the DOD level will be 3 percent lower than that at the start of this trip. Because it takes 3 hr for Line 3 to complete one round-trip, the hourly net battery DOD rate for Line 3 (defined as the difference in DOD levels between the end and the start of one round-trip, divided by the round-trip time) is 1 percent per hour.

The hourly net battery DOD rate for Bus Line 3 at 250 amp, together with the peak DOD reached during a round-trip, suggests that the battery could be discharged to the critical 80 percent level sometime during the 22nd round-trip, if the bus starts the day with a fully charged battery. At the start of the 22nd trip, the battery DOD would be approximately 63 percent. Adding the peak instantaneous DOD of 17 percent (which occurs at approximately Minute 77 of a round-trip) yields the 80 percent DOD value. This implies that an output current of 250 amp could allow Line 3 to operate for at least 21 round-trips (or a total of 63 hr) before the battery discharge would drop to 80 percent. Therefore, the ICS output current of 250 amp is more than adequate for the 15-hr daily range requirement of Line 3.

Strictly speaking, battery performance is not totally independent of the prevailing DOD value. As the battery DOD increases (i.e., the battery has less charge available), the battery can usually receive charge slightly more quickly. As a result, higher initial battery DOD values could yield slightly lower hourly net DOD rates. Therefore, the above projections of the daily range and ICS output current for Bus Line 3 (which assume that the average DOD at the start of all round-trips is about 20 percent) are accurate to the first order of approximation.

Examinations of the battery DOD characteristics for all nine bus lines indicate the following:

1. The ICS output current of 250 amp is adequate to ensure the 15-hr daily operation for Bus Lines 1, 2, 3, and 5.
2. Bus Lines 4, 6, 7, and 8 require the ICS output current of 300 amp to ensure the 15-hr daily operation.
3. Line 9, which is the longest bus line (i.e., 153 km for one round-trip, which takes almost 5 hr to complete), requires the ICS output current greater than 300 amp for 15-hr daily operation. However, a number of measures can be applied specifically to Line 9 so that it can operate for at least 15 hr per day with the ICS output current of 300 amp. These measures include (*a*) providing additional layover time at the existing layover points for more recharging time from static chargers and (*b*) identifying additional layover points along the very long suburban portion of the route for installations of additional static chargers.

The above analysis results suggest that Scenario A could be implemented with the ICS output current level of 300 amp.

Energy Consumption Under Scenario A

Net energy flow into the motor controller (i.e., energy consumption) under Scenario A is also determined from the EVSIM model. The results indicate that energy consumption for all bus lines does not differ much from one another and does not appear to be sensitive to the route length. Energy consumption values for all nine bus lines are found to range from 1.41 kW-hr/km (for Line 5) to 1.58 kW-hr/km (for Line 4).

Exclusive Use of Static Chargers on El Monte Busway

Because of the ease of implementation and lower infrastructure costs of static chargers (compared with the dynamic roadway electrification), it is of particular interest to determine whether static chargers can be used exclusively on El Monte Busway and in the downtown area without having to electrify the travel lanes of El Monte Busway. One scenario investigated is called Scenario D (or the downtown/El Monte shuttle bus service). Scenario D provides a shuttle bus service between the downtown area and El Monte Bus Terminal (Figures 2 and 3). The route for the downtown/El Monte shuttle bus essentially consists of two connecting loops: the downtown loop between Venice Boulevard and Union Station (Figure 3) and the El Monte loop between Union Station and El Monte Bus Terminal (Figure 2). The downtown loop is iden-

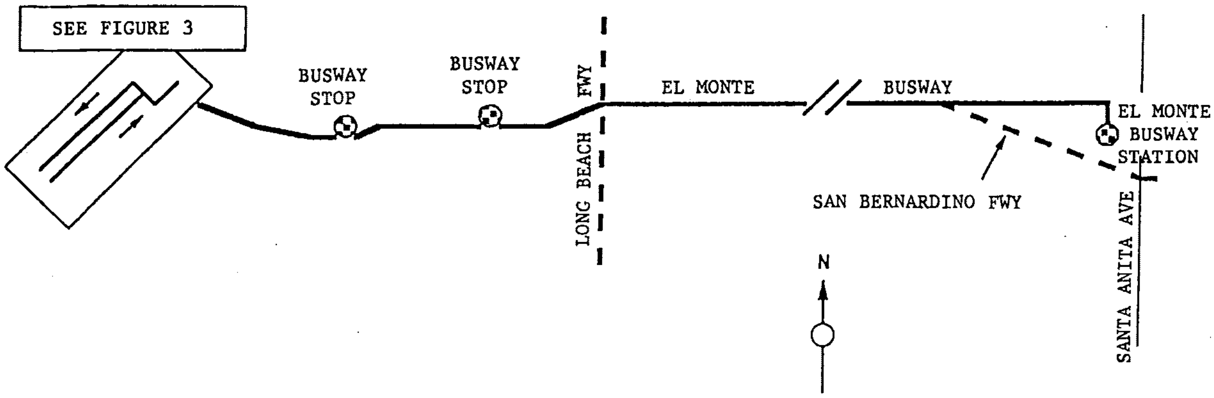


FIGURE 2 El Monte loop for Scenario D.

tical to the route of many Southern California Rapid Transit District buses and is similar to the existing downtown DASH bus service currently operated by Los Angeles Department of Transportation. The downtown/El Monte shuttle bus would start from Venice Boulevard, proceed onto Olive Street, First Street, Spring Street, and Aliso Street. It then lays over at Union Station before continuing onto El Monte Busway toward El Monte Bus Terminal. The shuttle bus, after the layover at El Monte Bus Terminal, would turn back toward Venice Boulevard.

Scenario D involves installing static chargers at downtown bus stops used by this shuttle bus as well as at three designated layover points (El Monte Bus Terminal, Union Station in downtown, and Venice Boulevard at the downtown end of the shuttle). Under Scenario D, the travel lane of El Monte Busway will not be electrified, which is why the infrastructure cost for Scenario D can be significantly reduced. The downtown/El Monte shuttle buses will be roadway-powered electric buses.

The role of static chargers used in Scenario D is to ensure that the shuttle bus can operate continuously for at least 15 hr/day without the battery DOD exceeding 80 percent. These static chargers are not meant to keep the on-board battery fully (or close to fully) charged at all times during the daily operation. Rather, by providing regular battery charging while the bus is in operation (i.e., short-duration charging at the bus stops and longer-duration charging at the layover points), it is possible to prevent the battery from being discharged below 80 percent during daily operation.

Scenario D offers a low-cost option for an early demonstration of RPEV technology. In particular, technology demonstration of the exclusive use of static chargers could precede the eventual and more expensive electrification of the travel lanes of El Monte Busway. In this way, real-world data could be collected to gain further understanding of various design aspects of RPEVs, such as impacts of vehicle loading and air conditioning on energy consumption, energy consumption under freeway and city street operating conditions, energy transfer characteristics, and battery performance and life.

Simulation using the EVSIM model is performed to determine whether Scenario D could provide adequate energy to enable the downtown/El Monte shuttle bus to operate con-

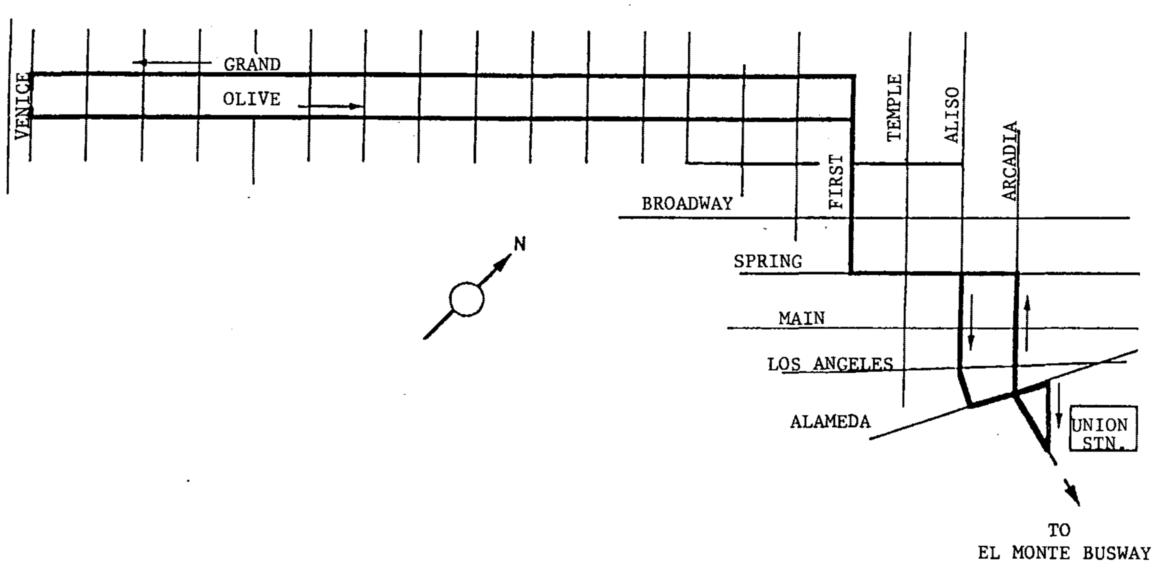


FIGURE 3 Downtown loop for Scenario D.

tinuously for at least 15 hr/day. The simulation results indicate that the total layover time (at the three layover points) in one round-trip affects the capability of Scenario D in supplying the adequate daily range for the shuttle bus. Longer layover time allows more time for battery recharging but increases the round-trip time. Higher round-trip time in turn implies that more buses will be needed to provide a particular service headway.

Figure 4 shows the hourly net battery DOD rate for Scenario D for four values of the round-trip time (75, 80, 85, and 90 min). Figure 4 indicates that the net battery DOD rate decreases with increasing values of the ICS output current up to 300 amp, beyond which no significant decline in this rate is expected. This implies that output currents greater than 300 amp are of no value for Scenario D. Figure 4 indicates that it is feasible to operate the downtown/El Monte shuttle service with the round-trip time of at least 80 min, with the output current of 275 to 300 amp, for 15 hr/day without the battery DOD exceeding 80 percent. For the round-trip time of 80 min, about 15 min would be layover time.

This paper also investigates a special limited case of Scenario D, in which shuttle buses could serve only the downtown loop between Union Station and Venice Boulevard (Figure 3) without serving El Monte Busway and El Monte Bus Terminal. This limited case will be called Scenario E (or the downtown shuttle bus service). Scenario E could have two layover points (at Union Station and Venice Boulevard) and involves only city streets. The simulation of the battery DOD for this downtown shuttle service indicates that it is feasible to use static chargers exclusively at these two layover points and at some bus stops along the downtown loop to provide adequate energy for the downtown shuttle bus to operate continuously for at least 15 hr/day, using the ICS output current of 300 amp. Specifically, the simulation indicates the following:

- Additional static chargers installed at all bus stops along the downtown loop (a total of about 38 stops) would enable the downtown shuttle bus to run indefinitely, if the round-trip time is at least 35 min, without the need for overnight battery recharging. For the round-trip time of 35 min, about 2 min is layover time.
- Additional static chargers installed at about 50 percent of the downtown bus stops (about 19 stops) would enable the downtown shuttle bus to operate for 15 hr/day with the round-trip time of at least 35 min, without the battery DOD exceeding 80 percent. However, the on-board battery has to be recharged overnight at the end of each daily operation.
- Additional static chargers installed at four to five downtown bus stops would enable the downtown shuttle bus to operate for 15 hr/day with the round-trip time of at least 40 min (about 7 min of which is layover time). Also, the battery has to be recharged overnight at the end of each daily operation.

Sharing of Power with Private Vehicles

The simulation results indicate that Scenarios A, D, and E are all viable in terms of providing adequate energy to satisfy the daily range requirement for transit buses. Under Scenario A, roadway-powered electric cars, vans, pickup, and so on, in addition to buses, can also draw power from the electrified travel lanes of El Monte Busway. On the other hand, Scenarios D and E are targeted to serve roadway-powered electric buses, with virtually no practical use for other vehicle types. This is because roadway-powered electric private passenger vehicles would not want to make frequent stops at bus stops or layover points to charge the batteries. From the bus operation's standpoint, it is also undesirable for the buses to share stops and layover points with private passenger vehicles.

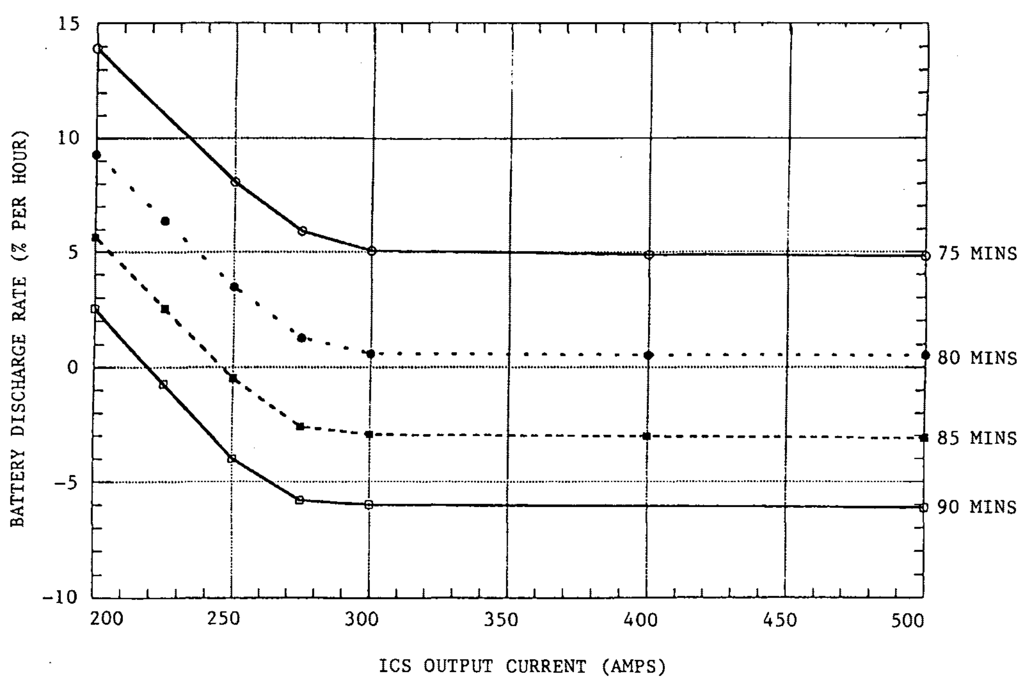


FIGURE 4 Hourly average net battery DOD versus output current (Scenario D).

ICS DESIGN FOR EL MONTE SYSTEM

Figure 5 shows the cross section of an ICS for the El Monte system, which consists of the roadway core and the vehicle pickup core. This ICS would operate at a power frequency of 8,500 Hz. Nominal operating point of this ICS has a roadway excitation of 250 amp-turns. This yields output current of 150 amp for one pickup (assuming a 7.5-cm air gap). Therefore, two pickups in parallel will provide an output current of (2 × 150) or 300 amp, as needed.

Roadway Cores

The roadway core module will be cast as a single W core, with a single multiple-conductor cable in each conduit. The roadway core module is made by embedding the core pieces, along with the conductor conduits, in a concrete matrix. A layer of fiberglass cloth and epoxy or polyester resin is applied over the cores.

Vehicle Pickup

The pickup cores will be the standard "hat section" type, made from hydraulically formed and oven-annealed laminations. For transit buses, split pickups with two 3-m-long sections would be used. Pickups for automobiles could be 1.5 m long. Flat-bar aluminum conductors are used for the conductor packs, which are encapsulated in fiberglass-reinforced epoxy to form a structural beam.

Power Conditioner and Distribution System

The power distribution system carries power from the power conditioners to individual roadway inductor segments. These segments are energized individually and activated only when an RPEV is present. The power distribution system senses the presence of vehicles equipped with the ICS and switches the segment on and off accordingly. Each power conditioner for the El Monte system can power 0.5 mi of travel lanes in both directions.

TECHNOLOGY DEMONSTRATION PLAN

A plan for the technology demonstration of the RPEV consists of two incremental phases, as described below and summarized in Table 2.

Phase 1: Implement Static Chargers Exclusively

The technology demonstration could start with the downtown/El Monte shuttle bus system (Scenario D) or the limited downtown shuttle bus system (Scenario E).

Selection of Scenario D

If Scenario D is selected, the route for the downtown/El Monte shuttle service is as shown in Figures 2 and 3. Static chargers would be required at the three bus layover points shown in the figures as well as at all downtown bus stops that lie along the downtown loop. Altogether, about 41 static chargers would be required. The round-trip distance for the downtown/El Monte shuttle service is about 43.7 km. This shuttle service could operate with total round-trip time of 80 min (of which about 15 min are layover time at the three layover points). To provide 10-min service headway, at least eight roadway-powered electric buses will be required.

Selection of Scenario E

If instead, Scenario E is selected for the demonstration in Phase 1, the route for the downtown shuttle service would involve only downtown streets and two layover points (at Union Station and Venice Boulevard), as shown in Figure 3. The round-trip distance for this shuttle service is about 8.5 km. Static chargers would be installed at the two bus layover

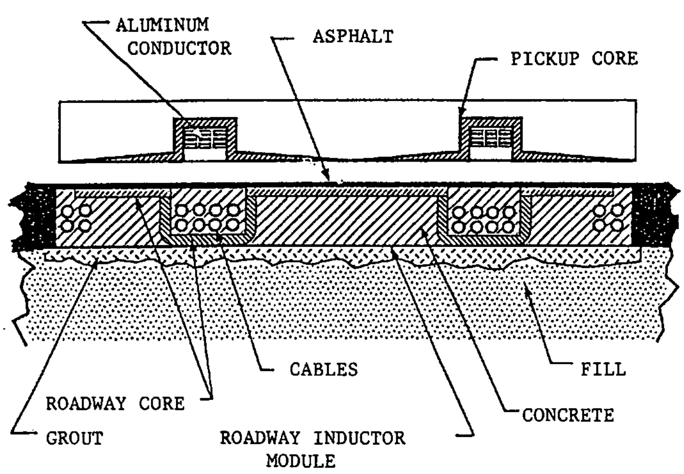


FIGURE 5 Cross section of ICS for El Monte system.

TABLE 2 Technology Demonstration Plan and Hardware Costs

Phase	Bus Headway (mins)	Round-Trip Distance (kms)	# Bus Required	# Static Chargers at Layover	#Static Chargers at Bus Stops	Kms. of Electrified Roadway	Hardware Cost ($ million)
Phase 1:							
o Downtown/El-Monte Shuttle Service	10	43.7	8	3	38	none	6.3
or							
o Downtown Shuttle Service	5	8.5	8	2	5	none	4.3
Phase 2: Electricity El-Monte Busway; operation of 9 bus lines	30*	**	68	12	38	about 36	91.8***

* The bus lines have average service headway of 30 minutes.

** Round-trip distance depends on the bus line.

*** This is an incremental amount over the $6.3 million spent in Phase 1.

points as well as at four to five bus stops along the downtown loop. The downtown shuttle bus could operate with the round-trip time of 40 min (about 7 min of which are layover time). To provide 5-min service headway, at least eight roadway-powered electric buses will be required.

Phase 2: Electrifying Travel Lanes of El Monte Busway

In Phase 2 of the demonstration, Scenario A could be implemented. This would involve electrifying about 36 lane-km of the travel lanes of El Monte Busway, as well as installing static chargers at the layover points of each of the nine bus lines and downtown bus stops. The static chargers previously installed in Phase 1 will become an integral part of the Phase 2 system. This scenario would enable all bus lines operating on El Monte Busway now to maintain their current routes and schedules; roadway-powered electric buses would simply replace existing diesel buses. It is estimated that under Scenario A, at least 68 roadway-powered electric buses would be required to provide service headway of 30 min.

Projected Hardware Costs for Technology Demonstration Plan

The hardware cost for RPEV systems includes the roadway subsystem (roadway inductor, power conditioners, and distribution system), engineering, roadway installation, vehicle, and vehicle equipment (pickup core and inductor, on-board power electronics, battery, on-board controls, and low-power vehicle steering system). There are uncertainties in estimating the hardware cost because the RPEV technology has not yet advanced to a product-development stage. Estimated hardware costs presented in this paper should be considered as approximations.

Because hardware costs for large-scale production are likely to differ from those for initial deployment, projected unit costs shown below are assumed to vary by the quantity used, as follows:

Item	*Cost*
Static chargers	$80,000 per charger for the first four, $60,000 per charger for the remainder
Electrified road	$2.5 million per lane-km for the first four, $1.9 million per lane-km for the remainder
Buses	$600,000 for the first bus, $450,000 per bus for the next 10, and $350,000 per bus for the remainder

Hardware costs for the two phases are shown in Table 2 (last column). Hardware costs for Phase 1 are estimated to be $6.3 million if the downtown/El Monte shuttle service (Scenario D) is selected for the demonstration. If the downtown shuttle service (Scenario E) is selected instead, hardware costs for Phase 1 would be lower ($4.3 million). This is because fewer static chargers are required for Scenario E than for Scenario D.

Estimated hardware costs for Phase 2 represent an incremental amount over the costs for the downtown/El Monte shuttle service. Estimated incremental hardware costs for Phase 2 are $91.8 million. Of this amount, about 76 percent is the cost of electrifying about 22 lane-mi of El Monte Busway, and the remaining 24 percent is the cost of acquiring 60 additional roadway-powered electric buses (eight buses already will have been acquired in Phase 1, which will be available for use in Phase 2). The estimated $91.8 million does not take into account the future savings from not having to replace 68 existing diesel buses when their service lives expire.

Energy Cost for RPEV

The energy cost for internal combustion–engine vehicles includes gasoline and oil, whereas the energy cost for the RPEV is primarily the electricity needed for charging from the roadway and overnight battery recharging. The estimation of energy costs for roadway-powered electric buses, full-sized vans, and cars assumes the following:

1. Average electricity cost is 6 and 10 cents/kW-hr (*5*) for overnight battery charging and roadway charging, respectively.
2. Energy consumption values for buses, vans, and cars are 1.56, 0.31, and 0.13 kW-hr/km, respectively (*5*).
3. Buses draw 80 to 85 percent of energy from the electrified roadway and 15 to 20 percent from overnight recharging. Roadway-powered electric vans and cars draw 25 percent of energy from the electrified roadway and about 75 percent from overnight recharging.

Estimated energy costs for roadway-powered electric buses, full-sized vans, and cars are found to be 14.7, 2.2, and 0.9 cents per vehicle-km of travel, respectively.

IMPACTS OF THE EL MONTE SYSTEM

The proposed plan for early deployment of the RPEV in El Monte Busway is a limited use of the technology. Therefore, environmental and energy impacts resulting from this early deployment are likely to be negligible. Later, when the technology has achieved its critical mass, its environmental and energy impacts could become more substantial. Southern California Association of Governments (SCAG) (*6*) and Bresnock et al. (*7*) reported that the primary long-term benefit of large-scale use of the RPEV would be significant reductions in vehicle emissions and dependence on petroleum. Bresnock et al. (*7*) reported that by electrifying over 1600 km of roadways in the SCAG region (assuming that 15 percent of the total VMT was caused by RPEVs) by 2025, peak utility demand could increase by about 1 percent relative to the baseline peak demand in which RPEVs are not deployed. Bresnock et al. (*7*) reported that the magnetic flux density measured from RPEVs was significantly below the standards for EMF exposure set by the International Radiation Protection Association and the International Non-Ionizing Radiation Committee; a flux density outside the vehicle at 1 m above the electrified roadway was lower than that from electric shavers. Possible long-term health and biological impacts of EMF are subjects of many ongoing research efforts, but there has been no conclusive finding to date.

CONCLUSION

The preliminary engineering feasibility study of early deployment of RPEV indicates that it is feasible to deploy a system in El Monte Busway as the first step toward larger-scale implementation on the highway. To advance the technology from its current state to the proposed technology demonstration, a number of activities need to be accomplished. They include detailed system design and specifications for the scenario selected for the demonstration, development and testing of prototype hardware, construction of operational facility and vehicle fabrication, and system shakedown and testing.

ACKNOWLEDGMENT

This paper is a part of larger research concerning the feasibility of advanced-technology HOV systems, which was funded by the Federal Transit Administration and California Department of Transportation.

REFERENCES

1. K. Lashkari, S. E. Shladover, and E. H. Lechner. Inductive Power Transfer to an Electric Vehicle. *Proc., 8th International Electric Vehicle Symposium*, Washington D.C., Oct. 1986.
2. E. H Lechner and S. E. Shladover. The Roadway Powered Electric Vehicle, An All-Electric Hybrid System. *Proc., 8th International Electric Vehicle Symposium*, Washington D.C., Oct. 1986.
3. S. E. Shladover. The Roadway-Powered Electric Transit Vehicle—Progress and Prospect. In *Transportation Research Record 1155*, TRB, National Research Council, Washington, D.C., 1987.
4. *Santa Barbara Electric Bus Project: Prototype Development and Technology Program, Phase 3B, Final Report.* System Control Technology, 1984.
5. *Electric Bus Project—Parametric Study, Final Report.* System Control Technology, 1992.
6. *Highway Electrification and Automation Technologies—Impacts Analysis Project Phase III.* Final Report. Southern California Association of Governments, Los Angeles, 1992.
7. A. Bresnock, M. Miller, E. Lechner, and S. Shladover. Roadway Electrification: Regional Impacts Assessment. Presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., 1993.

The findings, conclusions, and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the sponsors.

Publication of this paper sponsored by Committee on Intelligent Vehicle Highway Systems.

Phased Implementation of Lateral Guidance Systems in High-Occupancy-Vehicle Lanes

T. CHIRA-CHAVALA AND WEI-BIN ZHANG

Phased implementation of advanced lateral guidance systems in high-occupancy-vehicle lanes with exclusive right-of-way, or transitways, is described. The three incremental systems are steering assistance information systems, partially automated lane-keeping systems, and fully automated lateral control systems. Capabilities, functional requirements, system structure, and components for these systems are presented. Experiments and tests conducted at the University of California's PATH program indicate that the discrete magnetic roadway reference system appears to be a possible technology for applying all three incremental systems in transitways. Potential safety and capacity benefits of the three incremental systems in transitways are quantified.

Knowing the vehicle lateral position is important when operating vehicles on roadways. At this time, this knowledge and the task of keeping vehicles in the travel lane are drivers' responsibilities. Emerging lateral guidance and control systems could help to maintain the vehicle position along the lane center more precisely and reliably than drivers can. Therefore, these systems provide a means for improving the highway safety, particularly when drivers are inattentive or are driving in poor visibility. Furthermore, it is possible that control of vehicle lateral displacement may lead to reductions in the lane width requirement without compromising traffic safety. If so, additional travel lanes could be created and added capacity could result within existing rights-of-way.

Large-scale implementation of lateral guidance and control systems to enhance both highway safety and capacity is a long-term goal, and its progress requires incremental technology development and implementation, initially on a limited scale in existing highway facilities. High-occupancy-vehicle (HOV) lanes, which are separated from the freeway main lanes by permanent barriers and have controlled access and egress, are considered to be good candidates for this purpose. Implementation in these HOV facilities, which are generally known as transitways, could be a stepping stone to the eventual networkwide implementation because

- This limited-scale deployment of the new technology is likely to be less complicated than the networkwide deployment.
- The barriers separating transitways from the main lanes and the transitways' controlled access and egress could help to ensure maximum safety of the new system without on-the-road experience.
- This limited-scale deployment allows collection and evaluation of data concerning on-the-road system performance as well as driver responses to the use of driver-assisted devices. It could help to stimulate technology improvement toward full maturity.

Lateral control technology is not new. In fact, it has been investigated by researchers since the 1950s (*1–8*). However, research and development (R&D) efforts to apply lateral guidance and control systems in the highway environment are still in the very early stages. This paper examines an implementation strategy, a system structure, and the functional requirements of lateral guidance and control systems that could be deployed in existing transitways. It also assesses potential traffic and safety impacts due to this deployment.

OBJECTIVE

The objectives of this paper are to

- Describe lateral guidance and control systems for incremental implementation, from the near term to the long term, in existing transitways; and
- Assess the safety and traffic impacts of these incremental systems.

TECHNOLOGY OVERVIEW

The literature on vehicle lateral guidance and control systems emphasizes systems that perform the lane-keeping function. Such systems require roadway reference and sensing, vehicle controller, and vehicle actuation technologies. An overview of these component technologies is presented.

Roadway Reference and Sensing System

The roadway reference and sensing system requires both roadway reference and in-vehicle sensing devices. The roadway reference provides information about the lane lines or road edges. Possible roadway references include painted lines, raised

T. Chira-Chavala, Institute of Transportation Studies, University of California, 109 McLaughlin Hall, Berkeley, Calif. 94720. W.-B. Zhang, PATH Program, Institute of Transportation Studies, University of California, Building 452, Richmond Field Station, 1301 South 46th Street, Richmond, Calif. 94804.

pavement markers, and wires or magnetic markers embedded in the pavement to transmit electromagnetic or magnetic signals. The in-vehicle sensing device consists of sensors to receive signals from the roadway reference. These signals are then converted into information about the vehicle's lateral position. Depending on the kind of roadway reference and sensing technology employed, additional information such as upcoming roadway geometry (known as preview information) could also be obtained.

Roadway reference and sensing systems could be based on a number of technologies—for example, electromagnetic, radar, acoustic, optical (including vision-based), and magnetic. To date, these technologies have been investigated primarily in simulations and laboratory experiments. The two most promising technologies for large-scale applications in the highway environment appear to be vision-based systems and those using discrete magnetic markers (*9*).

Vision-Based Systems

Roadway reference and sensing systems based on vision-sensing technologies (e.g., video cameras) could directly acquire information on both the vehicle lateral position and the upcoming roadway geometry from existing lane lines or roadway delineation. Therefore, they are vehicle "autonomous" systems, which do not require special installation of roadway reference. At this time, there are uncertainties about applying vision-based systems on the highway because of the possible effects of debris, adverse weather, and light condition on performance; the requirement of large amounts of data in order to provide real-time guidance; and unknown reliability and costs of the in-vehicle data processing capability (*9*). R&D efforts to develop vision-based systems and their information-processing capability have been reported in the literature (*10–12*). However, whether these systems would work in diverse real-world weather and operating conditions remains largely unknown.

Discrete Magnetic Marker Systems

These roadway reference systems use small magnetic markers buried vertically in the center of the lane to provide the roadway reference. Magnetic fields from these markers are picked up by magnetic-type sensors installed on the vehicle. Relative to vision-based systems, discrete magnetic reference systems are not sensitive to debris or weather and light conditions (*9*). They could also provide preview information about upcoming road geometry by means of some magnetic coding scheme. However, they would require special installation of magnets in the pavement.

Other features of the discrete magnetic reference system include (a) the roadway in which the magnets are installed does not require electrification; (b) any magnetic damage or fault would affect the system locally but not the entire network, and repairs or replacements of the damaged magnets could be done quickly; and (c) the system could offer flexibility in that it is possible to install temporary markers to lead vehicles around construction zones, even without having to remove the markers already in place (*9*).

This paper focuses on applications of lateral guidance and control systems that use the discrete magnetic reference system in existing transitways. Research on such systems has been ongoing at the University of California's PATH program since 1987.

Vehicle Controller

Vehicle controllers for lane-keeping systems usually consist of computers and control algorithms. The computers process sensory information and execute control algorithms in real time. The control algorithms may include "feedback" and "feedforward" components. The feedback control algorithm receives information about the vehicle status with respect to the lane center and corrects local lateral deviations of the vehicle in an incremental manner. The feedforward control algorithm receives information on upcoming road geometry and issues commands to the steering actuation unit, in preparation for negotiating the change in roadway geometry ahead.

Vehicle Actuation Unit

The vehicle actuation unit receives steering commands from the vehicle controller for use in turning the steerable wheels. There are no commercially produced actuators for lane-keeping systems at this time, and research on such steering actuators is sparsely reported.

PHASED IMPLEMENTATION OF LATERAL GUIDANCE AND CONTROL SYSTEMS IN TRANSITWAYS

A possible strategy for implementing lateral guidance and control systems that use the discrete magnetic reference technology in existing transitway facilities is investigated. Implementation of these systems could follow a building-block approach, initially starting with systems that are relatively limited in terms of the degree of driver-assisted tasks. Three implementation phases are suggested; each is to be built on the preceding phase as far as possible. The three incremental systems are

- Phase 1: Implementation of steering assistance information systems (SAISs), essentially lateral warning systems, to enhance driver perception and warn drivers when vehicles are unintentionally encroaching on the lanes or drifting outside the lane; drivers would still perform all steering-related tasks themselves.
- Phase 2: Implementation of partially automated lane-keeping systems (ALKSs) to control the vehicle position along the lane center, with manual override for lane changes.
- Phase 3: Implementation of fully automated lateral control systems (ALCSs) that automatically perform lane-keeping, lane changing, merging, and diverging; the system could take over lateral steering tasks from the driver.

PHASE 1: STEERING ASSISTANCE INFORMATION SYSTEMS (SAISs)

Capabilities, functional requirements, system structure, and components of SAISs (that use the discrete magnetic reference and sensing system) to be implemented in transitways in Phase 1 are described. In addition, potential traffic and safety impacts due to the implementation of SAISs are also presented.

Capabilities of SAISs

It is foreseen that adoption of SAISs by transitway users will be voluntary. SAISs could provide the following real-time information to drivers:

- Vehicle lateral position: SAISs could provide information to drivers about vehicle lateral position with respect to the lane center. This is true in normal conditions, with poor or invisible lane marking, in poor-visibility conditions (e.g., nighttime), and in adverse weather.
- Edge warnings: SAISs could warn drivers when vehicles are inadvertently encroaching on adjacent lanes or drifting outside the lanes. These inadvertent vehicle maneuvers, which frequently result from driver fatigue, inattentiveness, or sleepiness, may be corrected by drivers if they receive warnings soon enough.
- Information about upcoming road geometry: SAISs could give drivers information about changes in the upcoming road geometry (e.g., road curvature). Therefore, they could help to better prepare drivers to make required steering actions sooner.

Functional Requirements

Functional requirements of new systems that have not yet been configured are defined as desired capabilities or performance goals that the systems should be targeted to achieve. Functional requirements for SAISs, with respect to the deployment in transitways, include

1. When an SAIS mistakenly gives a warning that the vehicle is drifting outside the lane when it actually is not, this warning is said to be a false alarm. Although false alarms per se may not be hazardous, a high rate of false alarms could erode user confidence. Effects of the rate of false alarms, and thus an acceptable false-alarm rate, for SAISs have not been explored. This knowledge is needed before reasonable rates of false alarms can be specified for practical SAISs. Nevertheless, it is believed that a very low rate of false alarms may be essential for public acceptance of the device.
2. When an SAIS fails to detect the vehicle drifting out of the lane, it is said to have a system "miss." System misses could be hazardous enough to result in traffic accidents. Therefore, system misses should be eliminated through the system design. Possible means for accomplishing this include the incorporation of sufficient redundancies for weak links within the system and the elimination of failures that could result in adverse consequences through systemic design.
3. Correct warnings on required steering correction actions (e.g., steer left or right) must be ensured. The incorporation of redundancies for weak links could help to enhance this system accuracy.
4. Warning signals to drivers must be effective, both when drivers are alert and when they are fatigued or inattentive.
5. The discrete magnetic reference and sensing system must be robust in all weather and operating conditions.
6. Installation and replacement of magnetic markers should be simple and it can be carried out quickly, because pavements are periodically resurfaced as part of regular maintenance. The magnetic markers should require minimal maintenance, and their life cycle should be compatible with that of the pavement.

System Structure and Components

A conceptual structure of near-term SAISs to be implemented in transitways is shown in Figure 1. Principal components of

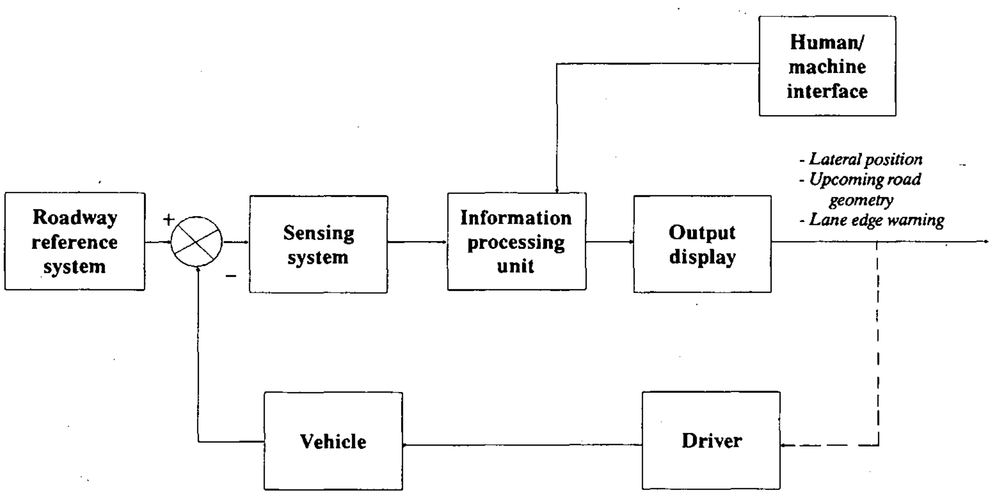


FIGURE 1 Steering assistance information system.

these SAISs include magnetic markers and an in-vehicle magnetic-sensing device (collectively called the magnetic roadway reference and sensing system), information processing unit, and human-machine interface unit.

Discrete Magnetic Reference and Sensing System

The discrete magnetic reference and sensing system that is being researched at PATH consists of a series of small permanent magnet markers. Each marker is 2.5 cm in diameter and 10 cm long and buried vertically in the pavement. These magnetic markers could be installed in single file along the lane center, with spacing of at least 100 cm. Dynamic tests on this system are being conducted at PATH to determine the effects of the magnetic-marker spacing on system performance. The magnetic fields provide information about the vehicle lateral position with respect to the lane center. In addition, by alternating the polarities of the magnetic markers, a series of binary information (0,1) can be encoded to provide information about upcoming roadway geometry. Codes could be checked and corrected to ensure the reliability of the preview information so obtained.

Magnetic signals generated from the magnetic markers are picked up by the on-vehicle sensing device. In this regard, magnetic sensors (or magnetometers) designed to be compatible with the magnetic markers can be installed under the front bumper of the vehicle. The signals acquired by the magnetometers are then converted into vehicle lateral deviation by the information processing unit. Tests of magnetometers conducted at PATH indicate that they are capable of acquiring signals at very low speeds (zero or close to zero) as well as at highway speeds (*5*). The magnetometers used in these tests are off-the-shelf devices. Magnetometers to be used by vehicles in transitways may have to be specially designed to ensure high reliability and low cost.

Tests were also conducted at PATH to measure signals from the magnetic reference and sensing system when pavements were covered with water and ice. Test results indicate that the signals are not degraded by these adverse conditions (*5*). Findings from experiments and tests conducted at PATH to date indicate that the discrete magnetic and sensing system appears to be a possible system for applications in transitways. Future tests of the discrete magnetic reference and sensing system will continue at PATH for roadways that have steel reinforcements.

Magnetic markers are relatively inexpensive. Those used in experiments at PATH are produced in small quantities and cost about $2 to $3 each. Production volumes of these magnets are likely to lower their cost. Magnetic markers are also relatively easy to install. Existing construction technologies (including the advanced robotics technology) could be adapted for the magnet-marker installation.

Information Processing Unit

The information processing unit consists of an onboard computer to process signals from the magnetometer and to produce output information for the drivers. An algorithm to process signals from the magnetometer has been developed at PATH (*5*). In addition, methods to overcome interference problems in the lateral position measurement—the overlapping with the earth's magnetic field, high-frequency magnetic noise generated by the vehicle engine system, and spontaneous vertical movements of the vehicle—were also developed and tested (*5*). This algorithm has since been tested at PATH, in bench tests, as well as in track tests using a scaled vehicle (1 m long and 0.5 m wide) and a full-size experimental vehicle. The test results indicate that this algorithm works satisfactorily.

Human-Machine Interface Unit

Information and warnings to drivers can be visual, audio, or both. For example, upcoming road curvature and the vehicle lateral position relative to the lane center could be displayed visually. Audio warnings could be given to the driver when the vehicle moves outside the lane. Warnings will not be given during deliberate lane-changing maneuvers. In this regard, the SAIS's on-off switch and the turn-signal switch could be integrated to deactivate the SAIS during a lane-changing maneuver and to resume the SAIS as soon as the lane change is complete. Research is needed to determine effective output display modes for practical SAISs.

Estimation of Potential Safety Benefit of SAISs

Implementation of SAISs in transitways is likely to have little direct effect on the transitway flow or capacity. However, it may bring about reductions in certain kinds of transitway accidents.

It is conceivable that SAISs could help to reduce frequencies of run-off-road and sideswipe accidents in transitways. This potential safety benefit is estimated using in-depth examinations of hard-copy accident reports as opposed to analyses of computerized accident data. This is because accidents on transitways could be difficult to identify from computerized data. Further, computerized accident data are not likely to have sufficient details for determining whether the accident outcome could be influenced by the use of new devices.

The transitway mileage in any one state is usually small. This plus the fact that accidents in transitways are even rarer events than accidents on the mainline make it necessary to obtain transitway accident data from a number of states in order to obtain a reasonable sample size. Hard-copy reports of transitway accidents from California, Houston (Texas), and Virginia (the I-66 transitway) are available for the analysis. The reports from California represent all reported transitway accidents for 4 months in 1990; the reports from Houston represent all reported transitway accidents for 12 months in 1990; and the reports from Virginia represent transitway accidents in the I-66 facility for 6 months in 1990. In all, 72 hard-copy reports of transitway accidents were analyzed.

For each commuter-lane accident, the in-depth analysis of the hard-copy accident report follows the following steps:

1. All information in the accident report, including the accident diagrams, is critically examined to identify a sequence of events and actions that culminated in the accident. This

sequence of events is useful for determining possible points of intervention by the new device.

2. Probable contributing factors of that accident are identified from all the evidence available in the accident report.

3. An evaluation is performed to see whether at least one of the identified contributing factors may respond to the new device. A view is taken that if one contributing factor could be eliminated by the new device, the new device would be considered useful as a possible accident countermeasure.

The results of this safety analysis indicate that SAISs could be useful as countermeasures for up to 8 percent of transitway accidents. These are accidents in which vehicles drift outside the travel lanes and strike the barriers or channelization at highway speed as a result of driver inattentiveness.

It must be noted that this estimated safety benefit is likely to be an upper-bound benefit for the following reasons:

1. Although it is foreseen that the adoption of SAISs will be on a voluntary basis for transitway users, the safety analysis assumes that all transitway users are equipped with SAISs. If SAISs are adopted by only a fraction of transitway users, the estimated benfit must be proportionally reduced.
2. In the accident analysis, it is assumed that SAISs will perform as they are expected to.
3. It is assumed that there would be no changes in driver behavior due to SAIS adoption.
4. The accident analysis cannot take into account the extent to which the use of SAISs may introduce new kinds of accidents, for example, those related to system malfunctions or failures. This determination requires on-the-road data that are not currently available.

PHASE 2: AUTOMATED LANE-KEEPING SYSTEMS

In Phase 2, automated lane-keeping systems (ALKSs) could be introduced in transitways. As with SAISs, the ALKSs investigated in this paper also use magnetic markers as the roadway reference system. Capabilities, functional requirements, system structure and components, and potential impacts of these ALKSs are described below.

Capabilities

ALKSs perform partially automated vehicle lateral control—that is, when the system is activated, it would automatically control the vehicle lateral position. However, this lane-keeping function could be temporarily deactivated when the driver engages the turn signal to perform manual lane-changing in transitways with two or more lanes. As soon as the manual lane change is complete, the automatic lane-keeping will be resumed.

It is expected that the adoption of ALKSs in Phase 2 by transitway users will be voluntary, and at least initially will not be accompanied by reductions in the lane width. Phase 2 is considered a desirable step before the implementation of fully automated lateral control systems because it would allow drivers to become familiar with using automated devices and learn to share tasks with them and it would allow on-the-road data concerning system performance to be collected for use in developing long-term fully automated lateral control systems.

Functional Requirements

In addition to the functional requirements previously described for SAISs, further functional requirements for the partially automated ALKSs include

1. ALKSs should be fail-safe systems. That is, system failures that could result in catastrophic consequences should be eliminated through the system design. Should system failures occur, they must not lead to a loss of vehicle controlability.
2. In this phase, drivers of equipped vehicles should have the option of turning the device on or off as they wish.
3. ALKSs must perform the lane-keeping task with good accuracy. In this regard, the allowable vehicle deviation (for ride quality, safety, and efficiency reasons) is being researched at PATH. Nevertheless, as an absolute minimum, ALKSs must be able to steer vehicles wholly within the lane boundaries.
4. ALKSs should have reasonably good ride quality in order to encourage system adoption. There are trade-offs between lane-keeping accuracy and ride quality that need to be addressed by further research.

System Structure and Components

ALKSs and SAISs share a number of components. Figure 2 shows a conceptual structure and major components of the ALKSs for Phase 2. The roadway reference and sensing system and information processing unit are identical to those for SAISs. In addition, ALKSs also require a battery of vehicle sensors, a vehicle control unit, and a steering actuator. It also requires a human-machine interface unit that is different from that for an SAIS. These additional components for ALKSs are described.

Vehicle Sensors

Vehicle sensors include accelerometers, angular-rate sensors, steering-angle sensors, and speed sensors for measuring vehicle lateral accelerations, yaw rates, actual ground wheel steering angles, and vehicle speeds, respectively. These technologies are largely available. However, more research is needed to assess whether their resolution and accuracy will be sufficient for applications in transitways.

Vehicle Control Unit

The vehicle control unit generates steering commands in accordance with some ride-quality and steering-accuracy requirements. This unit could share a computer with the information processing unit. Vehicle control algorithms would incorporate the "intelligence" that is capable of determining the vehicle status and environmental conditions for use in issuing steering commands approrpiate for prevailing condi-

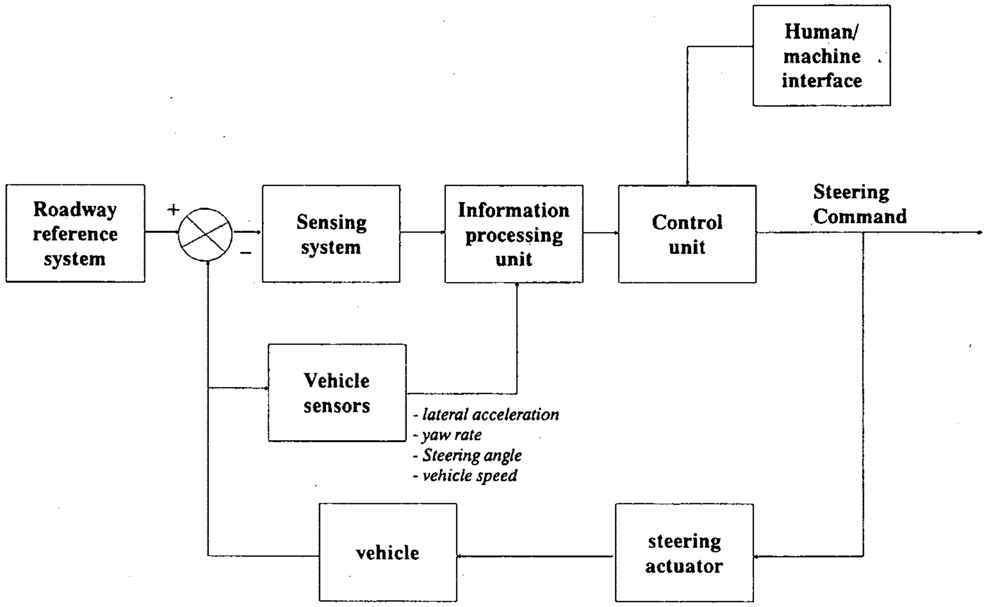


FIGURE 2 Automated lane-keeping system.

tions. The vehicle control unit could also incorporate safety logics to coordinate the transfer between automatic control and manual steering.

Both feedback and preview control algorithms have been developed at PATH (*7*, *8*). The feedback control algorithm generates steering commands from the feedback information (which includes the vehicle lateral position, lateral accelerations, and yaw). The preview control algorithm incorporates both the feedback and feedforward control components. The latter component estimates anticipatory steering angles from the information on upcoming road geometry.

Experiments on ALKSs were conducted at PATH (*13*). These experiment used a scaled vehicle (1 m long and 0.5 m wide), feedback and feedforward control algorithms, and the magnetic reference and sensing system. The test vehicle was equipped with electrical driving, a steering motor, a computer, and the aforementioned vehicle sensors. Information about upcoming roadway geometry was coded in the magnetic markers. The vehicle's maximum speed during the tests was 3 m/sec. In these tests, a maximum vehicle lateral displacement of ±20 cm was observed. Further, the test results also indicate that lateral accelerations could be controlled within an acceptable level required for good ride quality. Therefore, it appears that ALKSs using the discrete magnetic reference and sensing system are plausible systems for applications in transitways. PATH is planning to conduct further tests with a full-scale experimental vehicle. In this regard, a 700-m test track has been constructed at the University of California at Berkeley. These tests are expected to be completed in 1992.

Vehicle Actuation Unit

The steering actuator unit is used to operate steerable wheels to achieve required steering angles. These actuators, which may be hydraulic or electric servos, receive commands from the vehicle control unit. Research is needed to determine the maximum allowable steering angle for lane-keeping. One possible solution is to limit the maximum allowable steering angle of ALKSs to the minimum radius of curvature commonly recommended for the highway design.

Human-Machine Interface Unit

This unit is different than the one used in SAISs. For ALKSs, it is used to turn the automatic steering on and off. This interface unit can be designed to perform a number of functions. For example, the ALKS's on-off switch and the turn-signal switch could be integrated to facilitate manual lane-changing, as follows: this integrated switch could temporarily turn off the automatic lane-keeping when a lane-changing maneuver is taking place and resume the automatic lane-keeping once the maneuver has been completed. The interface unit could also incorporate a feature that would permit the driver to select ride-comfort levels versus tracking errors.

Research on human factors and the safety design of the human-machine interface for ALKSs is needed.

Impacts

The Phase 2 implementation of ALKSs could result in reductions in certain kinds of transitway accidents. Estimation of this safety benefit is presented.

Estimation of Safety Benefit of ALKSs

With the aid of ALKSs, the lateral position of equipped vehicles could be automatically controlled. This could eliminate driver errors of misjudgment in vehicle steering due to driver fatigue or inattentiveness; poor-visibility conditions; pavements covered with debris, water, mud, or snow; poor roadway delineation; or strong crosswinds. The previously mentioned in-depth analysis of hard-copy transitway accident reports indicate that ALKSs in Phase 2 could be useful as countermeasures for about 18 percent of transitway accidents, as follows:

- The 8 percent of transitway accidents for which the SAISs in Phase 1 are applicable as countermeasures;
- An additional 7 percent of transitway accidents that are run-off-road accidents on water-covered or icy pavements at highway speed, in which vehicles finally strike the barriers; for these accidents, the drivers did not state that they had actually applied brakes before running off the lane. It is not possible for the authors to determine, from the information available in the accident reports, how many of these accidents actually involved braking. ALKSs could be beneficial for those that do not involve braking, and they could also lower the probabilities of some accidents that involve braking.
- Another 3 percent of transitway accidents, which involve tire blowout that causes the vehicles to strike the barriers; that is, the probabilities of striking the barriers as a result of the tire blowout may be lower with ALKSs than without ALKSs.

The estimated accident benefit of ALKSs is likely to be an upper-bound benefit for the same reasons mentioned for the estimated safety benefit of SAISs.

Traffic Impact of ALKSs

The implementation of ALKSs in Phase 2, which is not accompanied by reductions in the lane width, is not expected to have significant direct impacts on the transitway flow rate or capacity. A possible exception might be the application in exclusive bus lanes (i.e., lanes specially reserved for buses) that have no shoulder or lane width smaller than 12 ft. If all the buses in the facilities are equipped with ALKSs, it is conceivable that ALKSs could help to counter the adverse effect, due to the lack of lateral clearance and narrow lanes, on the flow rate. ALKSs could eliminate the need for large lateral clearance between the vehicle and the roadside objects, which is deemed important for maximizing the flow rate under manual driving. The *Highway Capacity Manual* (HCM) documents a procedure for quantifying adverse effects due to the lack of lateral clearance on the flow rate (*14*). Based on this HCM's procedure, and if the ALKSs are assumed to be able to eliminate the need for full lateral clearance, the practical flow rate in single-lane bus lanes with no shoulder could increase by up to 13 percent. However, in the absence of actual on-the-road data concerning the use of ALKSs, human-factors research is required to verify this assumed benefit.

PHASE 3: FULLY AUTOMATED LATERAL CONTROL SYSTEMS

In Phase 3, long-term fully automated lateral control systems (ALCSs) could be introduced in transitways. Capabilities, functional requirements, system structure and components, and impacts of ALCSs are described.

Capabilities

ALCSs could take over the lateral steering of the vehicle. As with the Phase 2 ALKSs, ALCSs are capable of automated lane-keeping. In addition, ALCSs could also make other steering-related maneuvers, such as lane changes and merging. These additional automated capabilities call for the integration of additional devices. Further, all transitway vehicles would have to be equipped and the operating status of their ALCSs checked before entering the transitway to prevent failures due to equipment malfunctioning. With the mandatory system adoption by transitway users, significant lane-width reductions for the transitway would be possible without degrading traffic safety within the transitway.

Functional Requirements

Functional requirements for ALCSs are similar to those for the Phase 2 ALKSs, with the notable exceptions being the elimination of the manual override for lane changes and the driver option to turn the system on and off. From the safety perspective, such manual override and the on-off switch option appear to be undesirable for ALCSs. Further research is needed to determine if these features could be allowed in ALCSs.

System Structure and Components

Long-term ALCSs are the extension of ALKSs. Therefore, ALCSs would have all the components of the ALKSs plus some additional ones. As a minimum, the deployment of ALCSs in transitways would also require information links between individual transitway vehicles to facilitate automated lane-changing and merging maneuvers. These information links may be vehicle-to-vehicle or vehicle-and-roadside communication systems. Figure 3 shows a conceptual structure and components of the ALCSs.

Impacts

Primary impacts of the implementation of Phase 3 are reductions in frequencies of transitway accidents and increases in the transitway capacity. Other benefits include possible reductions in construction costs of future transitways, since smaller rights-of-way would be required. Also, when ALCSs are adopted, transitways can be constructed in locations where existing rights-of-way are currently not wide enough under existing transitway design guidelines.

Estimation of Safety Benefit of ALCSs

Similar to the ALKSs in Phase 2, ALCSs could eliminate driver error in lane-keeping, thus reducing accidents caused by such errors. In addition, the automated lane-changing and merging capabilities of ALCSs could also reduce the number of accidents related to lane-changing maneuvers in transitways. Results from the previously described in-depth analysis of

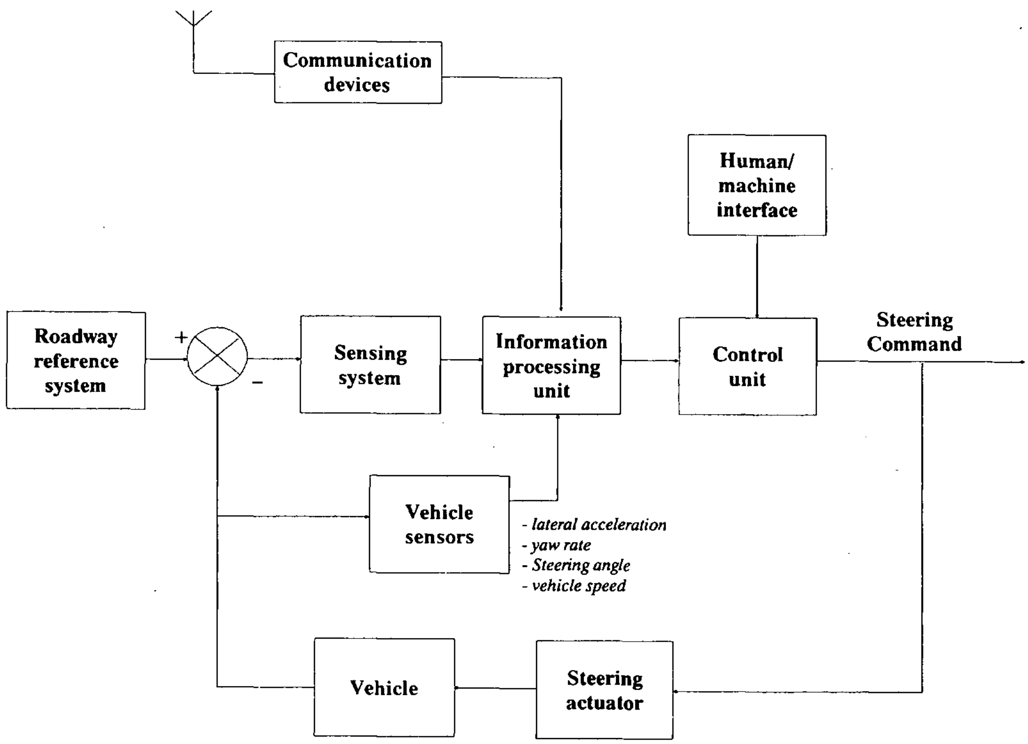


FIGURE 3 Fully automated lateral control system.

transitway hard-copy accident reports indicate that ALCSs could be useful as countermeasures for up to 24 percent of transitway accidents. Of these, 18 percent are the accidents for which the ALKSs in Phase 2 could be beneficial. The remaining 6 percent are accidents that occur during lane-changing maneuvers in transitways.

Estimation of Capacity Benefit of ALCSs

Adopting ALCSs could significantly increase transitway capacity if lane width could be reduced and additional travel lanes created within the existing right-of-way. This capacity increase is possible for existing transitways with at least two travel lanes. For most existing single-lane transitways, the lane-width reduction due to the use of ALCSs is not likely to be enough to create an additional travel lane. Nevertheless, small increases in the flow rate for single-lane transitways are possible if they currently have no shoulder or travel lanes less than 12 ft.

From the HCM's procedure for quantifying the effects of lateral clearance on the flow rate, possible increases in transitway capacity due to the Phase 3 implementation for various transitway configurations are estimated, as shown in Table 1. Table 1 also shows estimated changes in the transitway flow rate for the level-of-service (LOS) C as a result of adopting ALCSs. Estimates in Table 1 are based on the assumptions that transitway traffic is composed of 10 percent buses and 90 percent passenger vehicles and that the adoption of ALCSs could eliminate the need for the 12-ft lane width and full lateral clearance that are recommended in the HCM for flow-rate maximization under manual driving.

Table 1 indicates that, as a result of Phase 3 implementation,

- Practical capacity for single-lane transitways with no shoulder could increase by up to 14 percent.
- Practical capacity for existing two-lane transitways could increase much more substantially, when the use of ALCSs leads to reductions in the lane width and the creation of an additional lane (or lanes) within existing rights-of-way. Capacity increases of up to 47 percent and 60 percent could result for existing two-lane transitways with and without shoulders, respectively.

SUMMARY AND CONCLUSIONS

As with most intelligent vehicle-highway systems, lateral guidance systems are likely to provide the greatest safety and capacity benefits when implemented on most roadways in a network. However, networkwide implementation of fully automated lateral control systems is a very long term goal, and its progress requires incremental technology development and applications in existing highway facilities. Phased implementation of lateral guidance and control systems in existing transitways could be a stepping stone to networkwide implementation.

Phased implementation of such systems could begin with transitway users' adopting steering assistance information systems, which are essentially warning systems. Next, ALKSs

TABLE 1 Expected Capacity and Flow Rates at LOS C Before and After ALCS Adoption

HOV-Lane Configuration		Practical Capacity[e] (vph)	Flow Rate at LOS C[e] (vph)
BEFORE:	Existing 1-lane HOV facilities with no shoulder[a]	1,400	1,100
AFTER:	Same	1,600	1,300
BEFORE:	Existing 1-lane HOV facilities with shoulder[b]	1,600	1,300
AFTER:	Same	1,600	1,300
BEFORE:	Existing 2-lane HOV facilities with no shoulder[c]	3,500	2,600
AFTER:	3-lane HOV facilities with no shoulder[f]	5,600	4,500
BEFORE:	Existing 2-lane HOV facilities with shoulder[d]	3,800	3,200
AFTER:	3-lane HOV facilities with shoulder[f]	5,600	4,500

(a) 12-foot lane
(b) 12-foot lane, 5-8 foot lane
(c) 24-26 foot pavement
(d) 24-26 foot pavement, 5-8 shoulder
(e) Assuming 10 percent buses
(f) **New facilities with ALCSs have 3 lanes, two 8-foot lanes for automobiles and one 10-foot lane for buses plus automobiles**

could be introduced in transitways to provide automatic lane-keeping control with manual override for lane changes. Finally, ALCSs could be introduced in transitways to take over lateral steering tasks: lane-keeping, lane-changing, merging, and diverging.

Experiments and track tests conducted since 1987 at the University of California's PATH program on the discrete magnetic roadway reference indicate that it is a possible technology for applications of all three incremental lateral guidance and control systems in transitways.

Analyses performed in this paper indicate that the limited-scale implementation of SAISs and ALKSs in transitways could have safety benefits. The deployment of SAISs and ALKSs in transitways could reduce transitway accidents by as much as 8 and 18 percent, respectively. The implementation of ALCSs in transitways could reduce transitway accidents by as much as 24 percent. Further, if the lane width could be reduced as a result of adopting ALCSs, the capacity of two-lane transitways could be increased by 50 percent or more.

Even though the SAISs proposed in this paper are considered to be relatively near term systems for implementation in the highway environment, a significant amount of R&D remains to be completed or initiated before technology demonstration in transitways could take place. Principal R&D activities to be completed include

- Track tests of the SAISs using full-scale vehicles in real-world conditions;
- Tests to assess the compatibility between the discrete magnetic roadway reference and sensing system and existing roadway infrastructure, particularly the influence of steel reinforcements on the magnetic fields;
- Assessments of the cost-effectiveness of magnetometers; and
- The method in which warnings can be effectively and safely provided to drivers.

ACKNOWLEDGMENTS

This paper is part of a larger research effort conducted under the sponsorship of the California Department of Transportation and the Federal Transit Administration. The authors wish to acknowledge their sponsorship.

REFERENCES

1. J. W. Sanders, A. B. Kristofferson, W. H. Levison, C. W. Dietrich, and J. L. Ward. Attentional Demand of Automobile Driving. In *Highway Research Record 195*, HRB, National Research Council, Washington, D.C., 1967, pp. 15–33.
2. R. E. Fenton. Automatic Vehicle Guidance and Control—A State of the Art Survey. *IEEE Transactions on Vehicular Technology*, Vol. 10, No. 1, 1970, pp. 153–161.
3. D. T. McRuer, R. W. Allen, D. H. Weir, and R. H. Klein. New Results in Driver Steering Control Model. *Human Factors*, Vol. 19, No. 4, 1977, pp. 381–397.
4. S. E. Shladover, D. M. Wormley, H. H. Richardson, and R. Fish. Steering Controller Design for Automated Guideway Transit Vehicles. *Journals of Dynamic Systems, Measurement and Control*, Vol. 100, 1978, pp. 1–8.
5. W.-B. Zhang and R. Parsons. An Intelligent Roadway Reference System for Vehicle Lateral Guidance and Control. *Proc., 1990 American Control Conference*, San Diego, Calif., May 1990.
6. K. Ito, T. Fujishiro, K. Kanai, and Y. Ochi. Stability Analysis of Automatic Lateral Motion Controlled Vehicle with Four Wheel Steering System. *Proc., ACC Conference*, San Diego, Calif., 1990.
7. H. Peng and M. Tomizuka. *Lateral Control of Front-Wheel-Steering Rubber-Tire Vehicles.* Publication UCB-ITS-PRR-01-5. Institute of Transportation Studies, University of California, Berkeley, July 1990.
8. H. Peng and M. Tomizuka. Preview Control for Vehicle Lateral Guidance in Highway Automation. *Proc., 1991 American Control Conference*, Boston, Mass., 1991.
9. R. Parsons and W.-B. Zhang. PATH Lateral Guidance System Requirements Definition. PATH Research Report UCB-ITS-PRR-88-1. University of California, Berkeley, Oct. 1988.
10. H.-G. Metzer. Computer Vision Applied to Vehicle Operation. *Proc., SAE Conference on Future Transportation Technology*, San Francisco, Calif., 1988.

11. S. K. Kenue. *Lanelock: Detection of Lane Boundaries and Vehicle Tracking Using Image-Processing Techniques—Parts I & II.* General Motors Research Laboratories, Warren, Mich., 1989.
12. A. Hosaka, A. Takei, M. Taniguchi, and K. Kurami. The Development of Autonomous Controlled Vehicle, PVS. *Proc., Vehicle Navigation and Information Systems Conference*, Dearborn, Mich., Oct. 1991
13. T. Hessburg, H. Peng, M. Tomizuka, W.-B. Zhang, and E. Kamei. An Experimental Study on Lateral Control of a Vehicle. *Proc., 1991 American Control Conference*, Boston, Mass., 1991.
14. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1986.

The views expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

Publication of this paper sponsored by Committee on High-Occupancy Vehicle Systems.

Integrated Approach to Vehicle Routing and Congestion Prediction for Real-Time Driver Guidance

ISAM KAYSI, MOSHE BEN-AKIVA, AND HARIS KOUTSOPOULOS

The generation and dissemination of driver guidance that can be used for real-time diversion of traffic are expected to be implemented through the use of real-time traveler information systems. To implement these functions, a system structure consisting of a surveillance module, a congestion prediction module, and a control and routing (CAR) module is proposed, with the focus on the approaches that may be used for congestion prediction and the strategies that may form the basis for routing. It is argued that a congestion prediction capability is critical for the effectiveness of an on-line traveler information system. Such a capability is required to accurately forecast traffic conditions that may exist in the near future. The use of a dynamic traffic assignment model for congestion prediction is suggested. Such a model consists of dynamic driver behavior and network performance modules as well as origin-destination updating capability. Alternatively, statistical time-series methods may be necessary to generate predictions of future traffic conditions. The advantages and difficulties of adopting either approach are discussed. The predicted congestion information is passed to the CAR module to develop diversion strategies to alleviate both recurring and nonrecurring congestion. The role of routing strategies and update frequency in determining guidance effectiveness is discussed.

During the coming decades, efficient operation of existing road networks is expected to be achieved through dynamic traffic management schemes that make use of available and anticipated advanced technologies. Within this context, intelligent vehicle highway systems (IVHS) are currently being developed. These IVHS systems envision the linking of road infrastructure, vehicles, and drivers using advanced communication technology, computers, information display equipment, and traffic control systems.

In this context, advanced traveler information systems (ATIS) that are based on modern information technology may play an important role in reducing traffic congestion and improving traffic flows and safety. It is expected that ATIS will reduce delays caused by both incident and recurrent congestion by providing information to motorists about alternative paths to their destinations using a combination of roadside signals and onboard systems. Such schemes will aim at optimizing driver route selection and making this selection responsive to real-time road and traffic conditions. In this paper a framework for the operation of ATIS is presented and the modeling requirements of its constituent elements are analyzed.

I. Kaysi, Department of Civil Engineering, American University of Beirut, P.O. Box 11-0236, Beirut, Lebanon. M. Ben-Akiva and H. Koutsopoulos, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

INTEGRATED FRAMEWORK FOR IMPLEMENTATION OF ATIS

The actual benefits realized from traveler information systems depend heavily on the quality of the traffic information provided to drivers (*1*). This section describes the framework being proposed to provide drivers with guidance that they can have confidence in and that, as a result of improved information, can eliminate the occurrence of adverse impacts [see Ben-Akiva et al. (*2*)]. The discussion that follows describes the system structure and information flow embodied in the proposed framework. Later sections present the principles behind the proposed framework and describe the control and routing (CAR) and congestion prediction (COP) modules.

System Structure

A dynamic network modeling approach is critical to the effectiveness of real-time traveler information systems. Such a modeling approach is needed to accurately assess network performance as well as to forecast traffic conditions that may exist in the near future to develop real-time diversion strategies to alleviate both recurring and nonrecurring congestion conditions.

A proper framework for ATIS implementation should be able to integrate the functional needs referred to above into an operational system. Figure 1 illustrates the system structure and information flow of the framework within which real-time ATIS should be implemented (*3*). The functions performed by each element in the system are briefly described:

- The surveillance system consists of traffic sensors deployed on the various network elements (for example, detectors in the pavement, video cameras, possibly other optical recording equipment). The surveillance system may also include roadside readers that gather information about vehicles that are passing selected points on the network over time. Thus, equipped vehicles themselves may act as elements of the surveillance system by providing information on travel times on specific sections of the network. The collected data may consist of information on flows, speeds, travel times, the numbers of queued and moving vehicles on each link, and incident detection. The actual information gathered will vary from system to system, depending on the particular components included in the system and the coverage of the network by the surveillance system.

• The COP element has the responsibility for providing the control and routing module with the information that is needed to implement routing and guidance strategies. Among the principles adopted in this paper are the need for COP to provide CAR with projected traffic conditions and the fact that COP should be performed by a dynamic traffic assignment (DTA) model that will take into account in its projections the driver response.

• The CAR element generates guidance advice in response to information provided by the surveillance system and by the congestion prediction model. The fact that CAR should maintain projection or guidance consistency constitutes another principle adopted in this paper.

Information Flow

The flow of information from one element in the system to another is described as follows (refer to Figure 1):

• Infrastructure data: Infrastructure data, an umbrella term, includes all network attributes that generally are invariant with time or that change slowly with time. The attributes include the network topology, the geometric attributes of all network elements, the control devices that are installed in the network, any channelization or other type of lane control, circulation restrictions such as one-way streets and prohibited turns, and so on. This information is required by the COP and CAR.

• Historical origin-destination (O-D) data: The historical O-D data generally consist of O-D information obtained by surveys or inferred from traffic counts by assignment models.

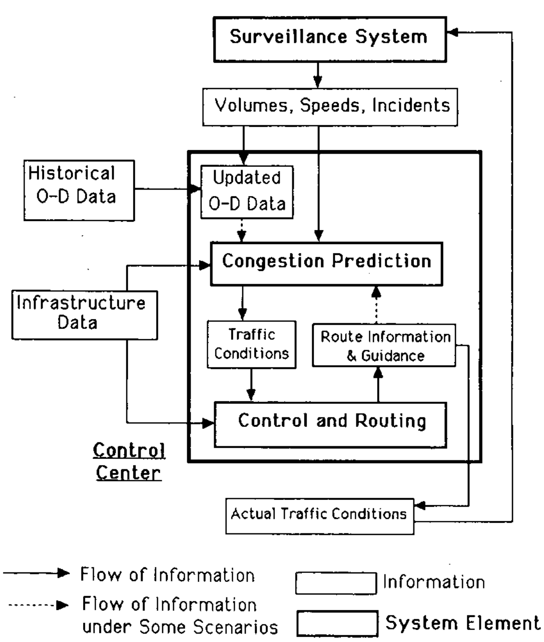


FIGURE 1 Proposed framework.

This historical information is, for the most part, slowly changing over time but should be updated periodically.

• The surveillance system: The information provided by the surveillance system consists mainly of direct measurements of volume, speed, occupancy, and the presence of incidents. Eventually the information may also include travel time data from equipped vehicles.

• Updated O-D data: The most recent information from the surveillance system and route information (if available) can be combined with historical O-D data to provide updated three-dimensional (3-D) O-D matrixes for the subsequent time periods.

• COP: Congestion prediction provides estimates of traffic conditions on the network. The updated O-D data are used by some COP scenarios including the proposed DTA. Because route choice modeling and the provision of guidance are sensitive to the destinations associated with flows, O-D data are required to implement the DTA associated with the proposed framework.

• CAR: The traffic conditions identified by the COP are transmitted to the CAR. The routing strategies need this information to develop an optimal response to the developing traffic environment.

• Guidance: The outputs of the CAR generally take the form of route guidance information. Within the proposed framework, guidance data are transmitted to the COP as an input.

PRINCIPLES UNDERLYING THE PROPOSED FRAMEWORK

The following are the major principles underlying the proposed framework:

Principle 1: COP Should Provide CAR with Projected Traffic Conditions

The travel times used for routing purposes by ATIS may be based on historical, current, or predicted traffic conditions. Although the use of historical data may provide a basis for static guidance and navigation, its use alone as a basis for CAR decisions is not expected to be of any value for adaptive routing. The main reason behind this is that historical traffic data are a bad indicator of evolving, day-specific traffic conditions, especially in situations in which traffic patterns display a significant amount of day-to-day variability. An analysis of the performance of various real-time routing strategies by Koutsopoulos and Xu (*4*) confirmed this intuition and indicated that the use of historical data as a basis for real-time routing advice is significantly inferior to the use of, for example, current or predictive information (*5*).

French (*6*), Catling and McQueen (*7*), and Rillings and Betsold (*8*) provide reviews of many existing demonstration projects. In many such projects guidance passed to drivers consists of information regarding current traffic conditions (Smart Corridor, AMTICS, and RACS). Some researchers assert that routing strategies may be formulated on the basis of a control-theoretic approach that requires information on current traffic conditions only. For example, Papageorgiou and Messmer (*9*) use feedback control methods to split traffic between an O-D pair among different routes. They claim that their methodology has low sensitivity with respect to unknown future demand levels and compliance rates that are assumed to be exogenous "disturbances." However, the authors warn that their feedback concept, which is based on observations of current traffic conditions, may not achieve its goal of establishing dynamic user optimum conditions if strong oscillations in the demand levels occur or if the network performance displays strong nonlinearities in case of severe congestion. These remarks by the authors provide further evidence that using current traffic conditions as a basis for guidance will not succeed if current traffic conditions are not good predictors of future conditions.

One of the major principles embodied in the framework being proposed in this paper is that such routing strategies have to be formulated on the basis of a forecast or projection of future traffic conditions on the network (*2*) rather than on instantaneous traffic conditions. The rationale behind using predictive information is that drivers' travel decisions are affected by future traffic conditions expected to be in effect when they reach downstream sections of the network on their way to their destinations. The ALI-SCOUT system uses projected travel times in setting the guidance on the basis of a similar rationale. Therefore, the most useful type of guidance that can be provided to a driver faced with travel decisions would be based on a projection of traffic conditions. In addition, guidance based on traffic information that is predicted using an advanced COP module is most capable of improving the travel time reliability of drivers because its look-ahead capability helps them avoid long future delays. This issue is discussed in more detail under the second principle.

Principle 2: A DTA Model Should Be Used for COP

In all existing traveler information systems and demonstration projects, the guidance passed to drivers is based either on current traffic conditions (Smart Corridor, Travtek, AMTICS, and RACS) or on simple predictions of future traffic conditions (ALI-SCOUT).

The travel time prediction methodology used by the ALI-SCOUT system, for example, constructs a projection ratio of the historical travel time on a specific link to the current travel time, as reported by equipped vehicles (*10*). This ratio is used to predict travel times for vehicles using that link during all future time intervals. Because only a few vehicles are equipped in Berlin, many links are not used by equipped vehicles during particular time intervals; therefore, no estimate of the projection ratio would be available. Consequently, the ratio is modified to reflect current conditions on neighboring links as well as conditions in preceding time intervals. Koutsopoulos and Xu (*4*) note that a problem with this methodology is the fact that the projection ratio is used to predict travel times for all future time intervals, thus implicitly assuming that trends currently observed remain constant for the entire prediction interval. To remedy this particular problem, Koutsopoulos and Xu suggest the use of information discounting. However, the methodology remains heuristic in nature and suffers from other omissions, which are discussed next.

In all these systems and projects, as well as in all analyses being conducted by researchers related to ATIS (*9,11*), there has been no consideration whatsoever of the response of motorists to route guidance in setting such guidance. Such an omission entails a major shortcoming in that the potential concentration of traffic on the recommended routes and the overreaction of drivers in their response to guidance information are ignored. This problem is expected to become more severe as the number of guided vehicles increases. Guidance validity and, as a result, driver compliance would be adversely affected in such schemes.

To overcome this problem it is required that the guidance be based on an advanced COP module that makes its predictions of future congestion in the network on the basis of

- Current traffic conditions (consideration of initial conditions),
- Predicted O-D demand levels (sensitivity to future demand patterns),
- Guidance being provided and anticipated driver response to guidance,
- Traffic control actions to be implemented, and
- Reduction in capacity as a result of incidents that have been detected.

Principle 3: CAR Should Maintain Guidance/ Prediction Consistency

On the basis of the earlier discussion, it becomes clear that none of the guidance systems in existence attempt to anticipate the impact of guidance being provided. The same holds true for analyses conducted by researchers in relation to ATIS. A major principle underlying the proposed framework is that consistency has to be maintained between the guidance being provided to drivers and the predicted traffic conditions. That is, the information system has to check that the guidance being provided will prove to be optimal to guided drivers on the basis of a prediction of the future traffic conditions. This system would result in guidance information that is consistent with anticipated driver behavior and network conditions, would ensure the validity of the guidance information, and would encourage its use by more drivers. In addition, consistency has to be maintained between the traffic conditions as perceived by the COP and the actual traffic conditions so that the COP remains attuned to the real world.

CAR STRATEGIES

When specifying the CAR module, three major issues have to be addressed:

1. What logic should the CAR module adopt to provide drivers with guidance advice? Should information or route directives be provided to drivers? Should route directives be based on shortest-path guidance or on route distributive guidance? How should route distributive guidance be implemented? How can the CAR logic ensure guidance and projection consistency?

2. Given that guidance provided at specific locations of the network can be updated periodically, what is the impact of temporal update frequency on the effectiveness of routing strategies?

3. If drivers receive guidance updates at various locations as they move through the network toward their destinations (as in ALI-SCOUT), what is the impact of spatial update frequency on the effectiveness of routing strategies?

A discussion of each of these issues follows.

CAR Logic

The most basic distinction between various logic types used as part of CAR relates to whether CAR provides drivers with information or route directives.

Information

If information is passed to drivers then CAR would constitute a simple link whereby the traffic data output from COP is interpreted, relayed to drivers, and presented in an understandable form. This represents the case in many demonstration projects, such as AMTICS, RACS, and Smart Corridor, and has served as the basis for ATIS analyses conducted by several researchers [see, for example, Mahmassani and Jayakrishman (*11*)].

Route Directives

Single Route Guidance Versus Route Distributive Guidance When route directives are employed, specific routing strategies may be based on the assignment of all traffic during a guidance interval to one path (usually the shortest path), whereas a different policy may be to distribute traffic over a number of alternate routes leading to the desired destination. Distributing traffic would entail the determination of the optimal fraction of traffic to be guided to the different routes, a process that may require significant computational effort in real time.

Jeffrey et al. (*12*) indicate their concern that guidance provided by AUTOGUIDE may overload a single "best" route to which drivers are guided. To deal with this problem they suggest that the "closeness" of journey times along alternate routes can be used to determine what proportion of vehicles should be sent along each route.

Papageorgiou and Messmer (*9*) used feedback regulation in an attempt to establish a dynamic user optimum whereby various routes used by flow departing at the same time between an O-D pair have the same travel time, and no unused path between the same O-D pair has a shorter travel time. The feedback regulator would observe current traffic conditions and determine the way traffic should be split among various paths connecting O-D pairs, with the aim of establishing the dynamic user optimum conditions. The first feedback law adopted consisted of shortest-route guidance. Vehicles traveling between a specific O-D pair were guided to the route currently having the shortest travel time for the O-D pair under consideration. However, the authors indicated that such logic may lead to strong perturbations of traffic flow, especially if a large fraction of vehicles are equipped. Therefore, they suggested an alternative logic based on a "smooth regulator" using a more advanced feedback law that leads to route-distributive guidance and the specification of optimal rates of splitting traffic among alternate routes.

Difficulties in Implementing Route Distributive Guidance Depending on the specific technology used for providing drivers with route directives, it may not be technically possible to distribute traffic over a number of routes at the same instant of time. For example, if variable message signs are used, the best that can be accomplished is to change the sign within the guidance interval so that the time average of the route directives would correspond to the fractions we wish to achieve for each route. On the other hand, if in-vehicle units are used, it is technically possible to provide different drivers with different route directives at the same time to split drivers among routes according to the optimal fractions. However, this process may not be politically or legally acceptable because of its inequity implications, and a scheme similar to what was suggested for variable message signs may have to be adopted.

In this paper it is proposed that shortest-route guidance be used whenever possible unless it does not succeed in maintaining projection/guidance consistency, in which case it becomes necessary to modify the CAR logic. This issue is discussed next.

Maintaining Guidance/Projection Consistency

It is important to recognize that there is a closed loop within the control center computations in the diagram of Figure 1. The loop involves the following: the COP sends data on projected traffic conditions to CAR, which in turn sends back to the COP information that defines the guidance environment that is in effect at specific times. This guidance information should be taken into account by the COP when projecting traffic conditions for the next forecasting period.

Specific route guidance should be provided only after going through the loop and having the CAR ensure that the guidance is consistent with traffic conditions projected by the COP. The guidance that is provided to drivers should represent a fixed-point solution of the COP/CAR interaction. For example, if shortest-path guidance is being used, consistency would be achieved if traffic conditions and travel times predicted by the COP indicate that vehicles are guided to the route that is predicted to be shortest.

If consistency is not achieved, the guidance advice has to be revised. Future traffic conditions resulting from the revised guidance have to be predicted by the COP and consistency with the provided guidance checked again. This loop may need to be repeated several times until consistency is achieved. If shortest-path guidance is being used then the scheme described above may not converge, and it may not be possible to achieve COP/CAR consistency. This is particularly true if the guidance is not updated frequently and therefore COP predictions take into account the response of a large number of drivers to guidance measures, with possible oscillations in potential messages and route use between subsequent executions of the COP/CAR cycle. Such oscillations and the implied lack of convergence would indicate that the control center is not capable of managing and mitigating the effects of the potential overreaction that will occur as a result of provision of guidance. This necessitates that the CAR logic using shortest-path guidance be modified to control the occurrence of overreaction and to deal with COP/CAR inconsistency.

Consistency Check Whenever the COP/CAR consistency check referred to above indicates a lack of convergence using single-route guidance, the travel time on the alternate routes would not be sufficiently different to warrant sending all guided traffic to the shorter route without causing overreaction. In such a case either the CAR logic should increase the temporal update frequency or guided vehicles should be distributed in such a way that they do not concentrate on any single one of the alternate routes.

Route Directives with Guidance Threshold The practice in existing route guidance experiments and studies of guidance being provided to a specific route whenever the travel time on such a route is evaluated to be shorter than that on alternate routes, even by a very small amount, often causes the guidance/projection inconsistency. As was indicated above, if all guided vehicles are sent to a route that is only marginally shorter than the alternate routes, overreaction is likely to occur. The use of a guidance threshold is suggested when route directives are in effect so that guidance is provided only if the travel time on the route to which drivers are directed is shorter than that on the alternate routes by a specific threshold. In the absence of a route that is shorter than the alternate routes by at least this threshold, guided vehicles have to be distributed over alternate routes. Such a scheme would have the potential to reduce overreaction associated with improved information and may be implemented on the basis of current or projected traffic conditions. Although it does not specifically identify occurrences of inconsistency as the "consistency check" scheme would, this scheme offers a computational advantage over that approach because a full projection/guidance consistency check (which requires a number of iterations and may involve checks over a long projection horizon) is not necessary.

Implementation In the case of collective route guidance technologies such as variable message signs the two schemes described above may implement "vehicle distribution" by simply providing "no guidance" and leaving drivers to distribute themselves as they would under normal traffic conditions. Such "natural" distribution is expected to yield good results because it would not impose excessive loads on any of the alternate routes. In addition, when it is called for, it would eliminate the computational and communication requirements associated with the determination of the optimal splitting fractions associated with a route distributive strategy, and the actual provision of such guidance.

In the case of in-vehicle units, on the other hand, drivers expect to receive route guidance at all times. One way of distributing guided vehicles in this case is to provide guidance in such a way that the time average of the directives relating to specific routes is equal to the average fractions of vehicles that would take each of these routes under normal traffic conditions. This would eliminate the computational effort required to determine the optimal splitting fractions, but guidance still needs to be communicated to the in-vehicle units.

Temporal Update Frequency

The frequency of guidance updates may influence the relative effectiveness of CAR schemes. If updating is infrequent, there is a real danger that a CAR logic such as shortest-path guidance may overload the shortest path and cause overreaction. In this case route distributive guidance might be necessary, at a possibly significant computational cost. On the other hand, if guidance is updated frequently, shortest-path guidance might perform satisfactorily and route distributive guidance might not be called for. The trade-off here is between the extra computational efforts involved in (a) computing the splitting ratios for distributive guidance and (b) providing more frequent updates in the case of single-route guidance. For instance, similar computational efforts may be required to provide distributive route guidance every 2 min or single-route guidance every 1 min for specific ATIS operating conditions. In that case, a comparison of the effectiveness of guidance provided in each of the two cases would be of interest. However, the specifics of this trade-off are not likely to become clear except after some experience with actual ATIS operations.

In addition, the guidance frequency may mitigate the impacts of using a less advanced information basis for CAR. For example, use of current traffic conditions as a basis for CAR would require very frequent updates of route guidance to avoid situations of serious overreaction in which congestion shifts from one location to another (*2*).

Spatial Update Frequency

The temporal update frequency discussed above relates to updates at a specific guidance location. Another consideration in the implementation of CAR strategies is the availability of spatial updates for a specific vehicle that is en route to its destination. Depending on the network structure, it may be possible for such a vehicle to receive guidance at more than one location during its trip. This would provide the CAR with possibilities of revising and correcting guidance that was provided to vehicles at an upstream location but that is no longer valid because of emerging traffic conditions (such as the occurrence of accidents). Such possibilities of spatial updates are potentially significant in improving the effectiveness of various CAR strategies. Koutsopoulos and Xu (*4*) observe that as the spatial update frequency increases, the adverse effects that were observed with low temporal update frequency at high guided fractions are somewhat alleviated. Clearly, this observation is consistent with our a priori expectations. A high degree of spatial update frequency is also likely to make an advanced COP module less necessary.

However, the benefits from a high spatial update frequency depend to a large degree on the network structure. Specifically, the network structure has to provide opportunities for reguidance and route diversion if the spatial update of guidance is to have any effect. For instance, for two routes in parallel with no crossovers, updating guidance along the route when it is not possible to switch to an alternate route would be worthless.

POSSIBLE COP SCHEMES

Types of Information Provided by COP to CAR

As discussed earlier, there exist two main types of information that could be provided by COP to CAR for use by routing strategies associated with ATIS. Some researchers hold the view that proper routing strategies require information on current traffic conditions only, whereas others assert that such routing strategies have to be formulated on the basis of a forecast or projection of future traffic conditions on the network.

COP Models

For the approach that bases its routing strategies on projected traffic conditions, there are the additional questions as to what projection methodology to adopt and the required accuracy of such methodology. These questions will be discussed next.

Conceptually, this scheme is more appealing although its implementation is more complicated. Two possible methodologies are presented below.

1. The first approach consists of a DTA model. This approach is described in more detail later. Briefly, the DTA provides the ability to project traffic conditions while taking into account potential driver response to guidance and the predicted time-dependent O-D matrixes. The DTA involves a prediction of driver behavior given the availability of guidance, an assignment of the predicted time-varying path flows to the network, and the subsequent determination of resulting flows and congestion on the various links of the network by time of day. To provide such capabilities, the DTA requires updated predictions of O-D demands, a dynamic driver behavior module, and a dynamic network performance module. The provision of guidance on the basis of projected traffic conditions is quite difficult in congested networks because such conditions are dependent on the ways in which drivers respond to the information. In other words, the validity of predicted network conditions depends on their consistency with current and future drivers' choices, which depend on the drivers' use of such information. This scheme would also carry with it a simultaneity problem to ensure consistency between provided guidance and projected traffic conditions.

2. Another approach to congestion prediction that involves significantly less computational and hardware requirements is the use of statistical time-series methods. Such an approach, which has been used for adaptive traffic control systems [see, for example, Stephanedes et al. (*13*) and Okutani and Stephanedes (*14*)], would make use of historical as well as recent traffic observations from the surveillance system to come up with congestion predictions. This approach is appealing because it does not require detailed modeling of O-D patterns, network structures, or driver response. However, its validity may be limited to relatively short projection periods. On the other hand, the behavioral models embedded in the DTA are more capable of capturing changes in traffic conditions over longer projection horizons. Both schemes are worthy of further investigation, and it might be that the best approach would be some combination of the two.

CONGESTION PREDICTION USING DTA

Proposed Approach and Elements of DTA

Some of the basic ideas related to the DTA approach proposed here were developed in previous research (*15–17*) that has been extended by Vythoulkas (*18*) to the case of a general network. A similar approach has also been suggested by Cascetta and Cantarella (*19*).

Because the DTA approach is proposed to serve as the COP module that will be used to predict the effects that travel decisions by informed drivers may have on overall traffic conditions, the approach should explicitly treat the distribution of traffic by time of day and drivers' pretrip and en route adjustments. This approach requires significant enhancements to the driver behavior and network performance models suggested in the research mentioned above.

The proposed DTA approach incorporates the following items:

- Dynamic driver behavior modeling,
- Dynamic network performance modeling, and
- O-D updating.

Dynamic Driver Behavior Modeling

A detailed discussion of dynamic driver behavior modeling for the DTA is not within the scope of this paper. At this point it is sufficient to say that the DTA approach has to capture the potential effects of the new information services on the departure time as well as pretrip and en route path choices of individual drivers. Therefore, models of the dynamic choices available to drivers with access to guidance should reflect the fashion in which new information concerning traffic conditions may affect driver behavior [see, for example, Kaysi (*20*) and Lotan (*21*)].

Dynamic Network Performance Modeling

The function of the DTA is to provide CAR with projected travel times on various links in the network. To perform this task, the DTA has to take into account the fact that traffic flows and network capacities vary with time. Temporal variations in traffic flows at various points of the network are caused by

- The time-varying nature of travel demand,
- The response of drivers to anticipated traffic congestion, and

- The delays at upstream bottlenecks that affect the arrival times at downstream facilities.

On the other hand, temporal variations in network capacities include primarily changes in capacities caused by incidents and traffic control actions.

Therefore, in a dynamic traffic assignment model all the following variables are space and time dependent: O-D trips, link capacities, link volumes, and link travel times.

Flow Computation Difficulties

Drivers' travel choices on a particular day (i.e., departure times and routes) translate into specific time-dependent path flows on the network. The dynamic network performance model of the DTA, whose elements are discussed below, has to determine time-dependent link flows and travel times that are consistent with predicted path flows. For simple networks (such as a number of routes in parallel between one O-D pair) this procedure is straightforward. However, for general networks, the correspondence between arc flows and path flows in a dynamic network is not as trivial as in the case of "steady-state" networks in which the two are simply related by a link/path incidence matrix. In dynamic networks, flow on any link during a specific interval is composed of path flows leaving their origin in that interval or in previous intervals and traversing (totally or partially) the link during the specified interval. Moreover, because of congestion, the time required to cross any link during a specific interval depends on the link flow during that interval and, for oversaturated links, on link flows in previous intervals.

Algorithms of varying complexity and levels of detail could be used to deal with this problem. These algorithms range from microsimulation in which individual cars are tracked through the network to macrosimulation algorithms that extend the familiar static network assignment algorithms to cope with time-varying demand and travel times. Many of the algorithms formulated to compute link flows from path flows in the case of within-day dynamic traffic conditions require the solution of a "fixed-point problem." Cascetta and Cantarella (*19*) solved the fixed-point problem in their DTA model by using meso-simulation, in which groups of vehicles are tracked through the network.

Dimensions of Dynamic Network Performance Models

The most important dimensions of network performance models consist of the time representation (continuous or discrete), flow representation (continuous flow, micro level, or macro level), and travel time computation. An analysis of the above items, including the proposed packet approach as well as an illustration of these ideas on a prototypical network and a discussion of the self-calibration aspect of the DTA, may be found elsewhere (*20*).

Updating the 3-D O-D Matrix

One set of inputs required by the dynamic traffic assignment models that are to be constructed are average (historical) O-D flows by time slice. These O-D matrixes differ significantly from the matrixes available for use in transportation planning analyses, for example. The required matrixes (representing dynamic flows) entail a much more refined description of time-varying travel demand between sufficiently small zones that represent the sources and sinks of the network. These dynamic O-D matrixes may be obtained from historically observed time-of-day flows using an off-line DTA.

Data that describe time-of-day variability in link flows or in O-D flows are almost nonexistent at this time. In other words, the 3-D O-D data required to implement the DTA are mostly unavailable, and it probably would take some time before a data base including the required type and level of detail of data were established. This process, however, is likely to be expedited by the substantial improvements in communication and computing devices and the resulting advances in traffic surveillance equipment.

Real-time traffic flow data obtained from the surveillance system are combined with the historical O-D matrix referred to above. The outcome of this process is an updated 3-D O-D matrix that is to be used in the next DTA projection period. The frequency of such updates will depend on the capabilities of the surveillance system to provide new extensive flow data as well as on the computational requirements of the updating process.

The problem of estimating O-D flows in real time on the basis of historical and recent measurements of traffic flows has received increased attention in recent years in view of its applicability for real-time traveler information and traffic control systems. The problem may be applied at the level of the network or individual junctions. Real-time estimation of O-D flows would require measurements of traffic flows at the entrances and exits of intersections. It is also useful to have measurements of travel times between counting locations. These measurements may obviously be obtained by the surveillance system. The aim is to estimate the proportion of traffic that goes from each origin to each destination.

DTA Approach

The dynamic O-D updating procedure would involve an iterative application of an off-line DTA. In this process a historical matrix would be adjusted to reflect the latest measurements from the surveillance system. Thus, the development of a real-time DTA would also benefit the process of dynamic O-D updating. This procedure is analogous to methods of static O-D estimation using a static equilibrium assignment.

Unfortunately, literature is not yet published on dynamic O-D estimation with a DTA. The existing literature on dynamic O-D estimation considers simple situations in which the distributions of travel times between counting stations are known. As such, in addition to unavailability of historical 3-D O-D data required to implement the DTA, the analytical tools required to provide predictions and updates of the O-D matrix have not yet been fully developed. The development of such tools will constitute part of the overall research effort required to implement ATIS.

Methodologies without a DTA that have been proposed to obtain the dynamic O-D estimates fall into two major categories: least-squares estimation and the Kalman filtering procedure.

Least-Squares Estimation

Variations on the least-squares approach (*22–24*) include constrained estimation whereby constraints are imposed on permissible values of O-D flows to ensure that fitted values of O-D flows are consistent with flow conservation and nonnegativity constraints; discounted estimation whereby recent observations are given more weight, a procedure that is necessary when the proportions of flow destined from a specific network entry point to various exits vary over time; and recursive least-squares estimation to improve the estimation efficiency.

Kalman Filtering Procedure

The Kalman filtering formulation is essentially a method developed for signal processing that provides optional estimates of the current state of a dynamic system. It consists of two equations: a transition equation and a measurement equation. The transition equation describes how the set of state variables (the parameters to be estimated) changes over time. The measurement equation provides predicted values of some observed variables on the basis of the (estimated) state variables. A comparison of the predicted and measured values of the observed variables provides a basis for updating estimates of the state variables for the next interval.

The application of the Kalman filtering approach to the problem at hand would have O-D flows represented as state variables that are to be estimated while the observed variables are the traffic counts. Note that this approach may be combined with a DTA as follows: dynamic traffic assignment, using the estimated O-D flows and playing the role of the measurement equation, will provide predicted values of traffic counts at specific stations. These predicted traffic counts are then compared with the actual observed counts at the stations. The prediction error detected from this comparison is used in the transition equation to estimate new values of the O-D flows for the next interval.

CONCLUDING REMARKS

This paper presented a possible system structure for the implementation of ATIS. The system consists of a surveillance system, a congestion prediction module, and a control and routing module. The paper illustrated the functional requirements of these modules and the interactions and information flows among them. The analysis indicated that models different from existing ones are needed to provide the functional capabilities of ATIS.

It was also observed that the occurrence and severity of overreaction may be reduced either by (a) basing the guidance on predictive traffic information that accounts for driver reaction, (b) providing guidance in such a way that traffic is distributed over several reasonable paths, or (c) performing more frequent guidance updates.

The analysis also identified several areas worthy of further research and investigation, including the formulation of more detailed specifications of the component modules and the possible need for algorithms for the implementation of optimal route distributive guidance. Finally, looking further ahead, prototypical tests have to be designed and conducted to examine the actual operation of the proposed system in the real world.

REFERENCES

1. P. W. Bonsall and A. D. May. Evaluation of In-Vehicle Route Guidance Systems. Presented at 5th WCTR, Yokohama, Japan, 1989.
2. M. Ben-Akiva, A. de Palma, and I. Kaysi. Dynamic Network Models and Driver Information Systems. In *Transportation Research*, Vol. 25A, No. 5, 1991, pp. 251–266.
3. M. Ben-Akiva and I. Kaysi. *Congestion Prediction Using Dynamic Traffic Assignment.* Working Paper. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Aug. 1991.
4. H. Koutsopoulos and H. Xu. An Information Discounting Routing Strategy for Advanced Traveller Information Systems (ATIS). *Transportation Research*, in press, 1992.
5. I. Kaysi, M. Ben-Akiva, and A. de Palma. Design Aspects of Advanced Traveller Information Systems. Presented at 2nd International Capri Seminar on Urban Traffic Networks, Capri, Italy, 1992.
6. R. L. French. Intelligent Vehicle/Highway Systems in Action. In *ITE Journal*, Nov. 1990, pp. 23–31.
7. I. Catling and B. McQueen. Road Transport Informatics in Europe—Major Programs and Demonstrations. In *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, 1991, pp. 132–140.
8. J. H. Rillings and R. J. Betsold. Advanced Driver Information Systems. In *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, 1991, pp. 31–40.
9. M. Papageorgiou and A. Messmer. Dynamic Network Traffic Assignment and Route Guidance Via Feedback Regulation. In *Transportation Research Record 1306*, TRB, National Research Council, Washington, D.C., Jan. 1991.
10. G. Hoffman and J. Janko. Travel Times as a Basic Part of the LISB Guidance Strategy. Presented at IEE Road Traffic Control Conference, London, 1990.
11. H. Mahmassani and R. Jayakrishnan. System Performance and User Response Under Real-Time Information in a Congested Traffic Corridor. Presented at Seminar on Urban Traffic Networks, Capri, Italy, June 1989.
12. D. J. Jeffrey, K. Russam, and D. I. Robertson. Electronic Route Guidance by AUTOGUIDE: The Research Background. In *Traffic Engineering and Control*, Oct. 1987, pp. 525–529.
13. Y. J. Stephanedes, P. G. Michalopoulos, and R. A. Plum. Improved Estimation of Traffic Flow for Real-Time Control. In *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981, pp. 28–39.
14. I. Okutani and Y. J. Stephanedes. Dynamic Prediction of Traffic Volume Through Kalman Filtering Theory. In *Transportation Research*, Vol. 18B, No. 1, 1984, pp. 1–11.
15. M. Ben-Akiva, M. Cyna, and A. de Palma. Dynamic Model of Peak Period Congestion. In *Transportation Research*, Vol. 18B, No. 4/5, 1984, pp. 339–355.
16. M. Ben-Akiva, A. de Palma, and P. Kanaroglou. Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates. In *Transportation Science*, Vol. 20, No. 2, 1986, pp. 164–181.
17. M. Ben-Akiva and A. de Palma. Dynamical Models of Transportation Networks. *Proc., PTRC Transport and Planning Summer Annual Meeting*, University of Bath, Bath, England, 1987.
18. P. Vythoulkas. A Dynamic Stochastic Assignment Model for the Analysis of General Networks. In *Transportation Research*, Vol. 24B, No. 6, 1990, pp. 453–469.

19. E. Cascetta and G. Cantarella. A Day-To-Day and Within-Day Dynamic Stochastic Assignment Model. In *Transportation Research*, Vol. 25A, No. 5, 1991, pp. 277–291.
20. I. Kaysi. *Framework and Models for the Provision of Real-Time Driver Information*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Feb. 1992.
21. T. Lotan. *Modelling Route Choice Behavior in the Presence of Information Using Concepts from Fuzzy Set Theory and Approximate Reasoning*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Aug. 1992.
22. M. Cremer and H. Keller. A New Class of Dynamic Methods for the Identification of Origin-Destination Flows. In *Transportation Research*, Vol. 21B, 1987, pp. 117–132.
23. N. Nihan and G. Davis. Recursive Estimation of Origin-Destination Matrices from Input/Output Counts. In *Transportation Research*, Vol. 21B, 1987, pp. 149–163.
24. A. Kessaci, J. Farges, and J. Henry. On-line Estimation of Turning Movements and Saturation Flows in PRODYN. Proc., *6th IFAC/IFIP/IFORS Symposium on Transportation*, Paris, France, 1989.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Experimental Analysis and Modeling of Sequential Route Choice Under an Advanced Traveler Information System in a Simplistic Traffic Network

KENNETH M. VAUGHN, MOHAMED A. ABDEL-ATY, RYUICHI KITAMURA, PAUL P. JOVANIS, HAI YANG, NEAL E. A. KROLL, ROBERT B. POST, AND BRIAN OPPY

An experiment to collect sequential route choice data under the influence of an advanced traveler information system was performed using a personal computer–based simulation. The experiment collected information on drivers' pretrip route choice behavior at three levels of information accuracy: 60, 75, and 90 percent. An analysis of variance was performed on the data to investigate the interrelationships among the different variables in an attempt to develop an understanding of what factors significantly influence route choice behavior and learning. An attempt was made to model sequential route choice behavior using a binary logit model formulation; the results were mixed. It was assumed that drivers update their knowledge of the system on the basis of their previous experiences; therefore an information updating function was specified and incorporated into the model. The results indicate that drivers can rapidly identify the accuracy level of information and that they adjust their behavior accordingly. Evidence also indicates that an accuracy threshold level exists below which drivers will not follow advice and above which drivers readily follow advice. It was found that male subjects agreed with advice more often than females, that less experienced drivers agreed more often than experienced drivers, and that a "freeway bias" exists with drivers much more willing to follow advice to take a freeway route. The model of route choice behavior had a prediction rate that was 79 percent accurate, which also indicated that previous experiences had little effect on current route choices. This value may be the result of a misspecified updating function, indicating that further research is required to identify these learning relationships.

The route choice process in the real traffic environment is very complex and there is little experimental evidence of how drivers process information and select their routes (*1*). Therefore, it was decided to analyze route choice behavior in the most simplistic, controlled environment possible. It was believed that this level of control would allow adequate restriction and analysis of the effects of various factors on the route choice behavior. The factor of utmost importance to any analysis of driver behavior influenced by an advanced traveler information system (ATIS) is a measure of the information accuracy. Certainly the future success or failure of ATIS will be highly dependent on the accuracy as well as the quality of advice that can be consistently delivered to the drivers. Previous research (*2–4*) has indicated the influence of system accuracy on compliance with advice. If a system consistently provides bad information it is assumed that drivers will soon begin to ignore the advice, and route choice patterns will remain unchanged. If highly accurate information is consistently provided to drivers it is assumed that drivers will perceive a benefit from following the advice and adapt their behavior to the advice. How do drivers perceive the accuracy of information? Is there an accuracy threshold below which drivers perceive no benefit from following advice? If such thresholds exist are they consistent for all drivers, or do different types of drivers have different thresholds? Can drivers perceive the accuracy of advice, under what conditions, and how rapidly? All of these questions need to be addressed to maximize the potential of ATIS.

The analysis suggests that initially drivers are predisposed to follow the route advice. The average agreement with advice over time shows that for the first few trials drivers accept the advice approximately 78 percent of the time independent of the accuracy level of advice being provided. The findings also suggest that drivers can perceive the level of information accuracy and that they do so rather rapidly. Within the first 8 of 32 sequential trials, the average agreement with advice moved in the direction of the level of accuracy provided. At 75 and 90 percent levels of accuracy, the average agreement with advice increased over the remaining 28 trials, whereas at the 60 percent level of accuracy, the average agreement declined from the initial rate to approximately 60 percent (system accuracy). These findings indicate the importance of the accuracy of information provided by ATIS and show that drivers can quickly discern the level of accuracy being provided.

DESCRIPTION OF ROUTE CHOICE EXPERIMENT

An experiment to investigate drivers' learning and pretrip route choice behavior under ATIS was performed using an

K. M. Vaughn, M. A. Abdel-Aty, R. Kitamura, P. P. Jovanis, and H. Yang, Department of Civil and Environmental Engineering; N. E. A. Kroll, R. B. Post, and B. Oppy, Department of Psychology, University of California at Davis, Davis, Calif. 95616.

interactive route choice simulation experiment carried out on a personal computer. The experiment was developed through a collaborative effort between the Institute of Transportation Studies and the Psychology Department at the University of California at Davis.

The simulation begins by presenting a set of instructions to the subject describing how the program operates. The subjects are told that they have purchased a new "traffic watch device" that will provide them with traffic information before they select a route. The subjects are also told that the device will not always be accurate but are not given any indication of its overall accuracy. Before beginning the simulation the subjects are shown examples of the fastest and slowest possible times on each of the routes, and they may repeat the examples as often as necessary to get familiar with the system. Subjects are instructed that their main task is to minimize their overall travel time by deciding when and when not to follow the advice provided by the traffic information system. Subjects are also told that their decisions and response times are being measured and that they should try to respond as quickly as they can make a good decision.

When the subjects are ready to begin the simulation, they are presented with a screen that indicates that it is Trial Day 1; they are instructed to position their hands on the computer keyboard and to press the space bar when they are ready to receive advice. On pressing the space bar, the advice for that day is presented along with a simulated freeway link, a side road link, and an origin and destination. The advice given was either, "Take the Freeway, traffic is moving smoothly" or "Take the side road, there is a problem on the Freeway." The screen display was simple and is approximated in Figure 1.

When the subject selects a route, a red blinking curser (shown by a shaded box on the freeway link) moves across the screen from the starting point (S) to the goal (G). The speed at which the curser moves represents the average travel speed on that link for that travel day. In Figure 1, the double line link represents the freeway and the single line link represents the side road. On completion of each trial subjects were asked to rate their choice satisfaction (e.g., correct, probably correct, don't know, probably incorrect, incorrect) and to provide an estimate of their travel time on their chosen route (e.g., fastest possible, reasonably fast, moderate, fairly slow, incredibly slow).

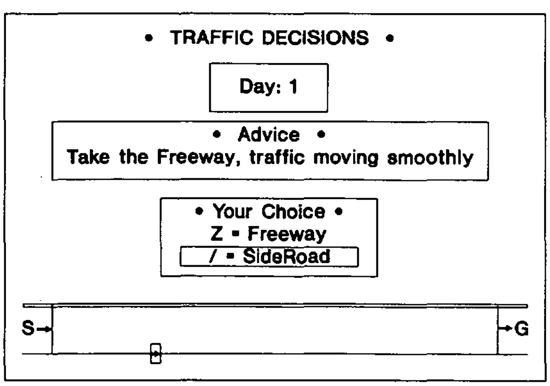


FIGURE 1 Typical screen display of simulator.

The simulation was developed such that various treatments could be applied and then data could be collected under these different conditions. The treatments that could be applied to the simulation included the following:

1. Accuracy: The accuracy level of the advice provided to subjects could take on values of 60, 75, or 90 percent.
2. Stops: A simulated stop on the side road route could be applied.
3. Rationale: A justification statement about why the subject should follow the advice could be provided.
4. Feedback: Feedback could be provided at the end of each trial in the form of actual simulated travel times on the two routes for that trial.
5. Freeway: Identification of the routes as freeway and side road as opposed to simply Routes A and B could be provided.
6. Road: The display could provide the simulated origin and destination with the two route links as shown in Figure 1 or with no network display provided and the travel time simulated by a blinking curser located in the center of the screen.

Three separate experiments were carried out to collect data under various conditions. The three experiments and the conditions under which the simulation has been run to date are shown in Table 1. The first experiment was used to investigate accuracy requirements of ATIS. The experiment was structured as described above but with three levels of information accuracy provided. Three separate groups of 23, 25, and 29 subjects were run through the simulation at three levels of accuracy: 60, 75, and 90 percent. In the second and third experiments the information accuracy was held constant at 75 percent while other experimental conditions were varied. This paper provides an initial analysis of the data collected in the first and third experiments using 4 of the 16 possible initial conditions (Conditions 1, 2, 3, and 7). A forthcoming paper will address the second and third experiments and the effects of varying conditions.

All of the experiments subjected drivers to 32 simulated days in which they were to choose one of two possible routes. For each travel day an amount of delay was randomly assigned to each of the two routes. The units of delay assigned to a particular route are proportional to the travel time experienced on the route. The delay was distributed over the 32 trials such that the mean delay for each route was equal but the variance differed. In this manner, routes with potentially faster travel times but with a greater amount of uncertainty (as one might expect on a freeway) can be compared with routes with slower travel times but with a greater amount of certainty (similar to surface street routes). On completion of 32 sequential simulated days, subjects were asked to rate their potential for purchasing a traffic information device, their perceived accuracy of the device, and their own ability at selecting routes when compared with that of the information device.

The computer program automatically recorded and stored data from each subject for 32 sequential trials. Test subjects were all undergraduate students in the psychology department at the University of California at Davis.

INVESTIGATION OF BEHAVIORAL RELATIONSHIPS USING ANALYSIS OF VARIANCE

Analysis of variance (ANOVA) models are used for studying the relationship between a dependent variable and one or more independent variables for experimental and observational data. The strength of the ANOVA model, and the main reason it is applied here, is that it requires neither that assumptions be made about the nature of the statistical relation nor that the independent variables be quantitative (*5*).

Fixed Effects Model

The goal of this research effort is to develop models of route choice under the influence of ATIS and to capture and incorporate into these models the effects of drivers' learning abilities. The first step in this process is to develop a basic understanding of the factors that influence driver's route choices and how the presence of traffic information systems will affect drivers' route choice decisions over time. The experiment described above was developed explicitly to study drivers' route choice behavior at its most basic level.

The first step in the data analysis was to investigate the interrelationships among the various variables in an attempt to develop an understanding of what factors significantly influence route choice behavior and learning. Three variables of significant interest were selected from the data set for analysis as dependent variables. The first variable of interest is one that indicates a driver's willingness to accept the route choice advice that is provided by the information system. This is a variable that compares the route choice made by Subject i on Travel Day j with the advised route for that day and returns a value of 1 if the subject chose the advised route and a value of 0 otherwise. This variable was analyzed in two different formulations—the first being the average acceptance rate of the advice given and the second being the individual agreement or disagreement. The average acceptance rate is the average acceptance of advice for Subject i on Travel Day j and is given by

$$\text{ACCRATE}_j = \left(\textstyle\sum_{l=1..j} \text{AGR}_l\right) / j \qquad (1)$$

where AGR_l is the level of agreement on day l (1 = agree, 0 = disagree).

The second dependent variable of interest included in the analysis is the subject's decision time. This is the time in seconds that a subject takes to select a route once route information has been provided. The third variable analyzed was a subject's potential use of an information system of the type experienced in the simulation. This variable is a subject's rating of how likely he or she would be to buy such an information system.

The data set used for the ANOVA consists of the 2,464 individual choices made by the 77 subjects from Experiment 1 (Conditions 1 through 3) as shown in Table 1. A software package was used (BMDP 2V) to perform an analysis of variance and covariance on fixed effects factorial designs with

TABLE 1 Experimental Treatments

Experiment	Condition	Treatment Number						Number of
		1	2	3	4	5	6	Subjects
1	1	60%	no	yes	yes	yes	yes	23
1	2	75%	no	yes	yes	yes	yes	25
1	3	90%	no	yes	yes	yes	yes	29
2	4	75%	yes	yes	yes	yes	yes	20
2	5	75%	yes	yes	yes	yes	yes	20
2	6	75%	yes	yes	no	yes	yes	20
3	7	75%	no	yes	yes	yes	yes	20
3	8	75%	no	no	yes	yes	yes	20
3	9	75%	no	yes	no	yes	yes	20
3	10	75%	no	yes	yes	no	yes	20
3	11	75%	no	no	yes	no	yes	20
3	12	75%	no	no	no	no	yes	20
3	13	75%	no	no	yes	no	no	20
3	14	75%	no	no	no	no	no	20
3	15	75%	yes	no	yes	no	no	20
3	16	75%	yes	no	no	no	no	20

two grouping factors (two-way ANOVA). For two-way ANOVA, tests are made of the null hypotheses about equality of main effects for each factor and about interactions between factors. Five variables were selected as grouping variables, and five variables were selected as covariates with three of the variables overlapping. The grouping variables included the percent accuracy of advice (PAADV), the trial block of the choice (TBLOCK), the driving frequency of the subject (NDFREQ), the advised route (ADVROUTE), and the subject's gender (SEX). The covariate or independent variables included SEX, NDFREQ, the individual trial number (TRIAL#), accuracy on previous trial (ADV__N-1), and ADVROUTE. The ANOVA model used in this analysis is the factor effects model for two-way factor studies (*5*). It was decided to use a two-way factor study and include covariate terms as opposed to performing a full multifactor study, which was not feasible because of the relatively small sample size.

The findings indicate that the willingness of subjects to follow advice is strongly influenced by the accuracy of the advice, the experience level of the driver, the gender, and the route being advised. The effects of learning, gauged by the number of trials, were shown to have little effect on the willingness to follow advice. Subjects' decision times were significantly influenced by the accuracy of advice, gender, the advised route, driving frequency, and system experience (trial block). The potential use of an information system was shown to be influenced by the accuracy of advice, the gender, and the driving frequency.

Regression Model

The constant vectors of the ANOVA factor effects model give an indication of the effects of the within-factor levels of the grouping variables on the dependent variable. These factor level constants can be estimated using a regression approach that is equivalent to the ANOVA model (*5*). An in-depth description of the procedures and results of the ANOVA is given elsewhere (*6*).

From the ANOVA, the factors and covariates that had significant effects on the dependent variables were determined. For the dependent variable ACCRATE, the grouping factors PAADV and SEX and the covariate NDFREQ are the most significant. For the dependent variable AGR, the grouping factors PAADV and ADVROUTE and the covariates SEX and NDFREQ are the most significant. For the dependent variable DTIME, the grouping factors PAADV and ADVROUTE and the covariates SEX, NDFREQ, and TRIAL# are the most significant. For the dependent variable USAGE, the grouping factors PAADV and SEX and the covariate NDFREQ are the most significant. For each of these dependent variables a regression analysis was performed to determine the within-factor coefficients of the regression equation. The results of the regression analysis for these four models are presented in Table 2.

The ANOVA results gave an indication of which variables have significant effects on subjects' willingness to follow route advice, their decision times, and their potential use of an information system. The ANOVA regression technique provided insight into how these variables influence route choice decisions. The significant findings of this section are summarized below:

1. Acceptance of advice increases with increasing information accuracy.
2. Males are more willing to accept advice than females and also make their decisions faster than females.

TABLE 2 ANOVA Regression Coefficients

	DEPENDENT= AVERAGE ACCEPTANCE		DEPENDENT= AGREEMENT		DEPENDENT= DECISION TIME		DEPENDENT = USAGE	
INDEPENDENT	COEFF.	T	COEFF.	T	COEFF.	T	COEFF.	T
INTERCEPT	0.78867		0.71504		3.36649		2.7493	
SEX (1=F, 2=M)			0.0382	2.21	-0.1431	-1.87		
DRIVING FREQ. (1=HI, 2=MED, 3=LOW)	0.0115	2.81	0.0195	1.95	0.1905	4.3	0.1532	5.15
TRIAL# (1 - 32)					-0.0768	-19.55		
X1 (65% accuracy)	-0.043	-9.25	-0.0926	-7.99	-0.0787	-1.53	0.6629	19.58
X2 (75% accuracy)	-0.0125	-2.65	0.0033	0.29	0.2239	4.36	0.0447	1.30
X3 (side road advice)	-0.0375	-11.42	-0.0685	-8.36	0.4029	11.11	-0.0879	-3.67
X13 (X1 X3 interactions)	-0.0064	-1.34	-0.0282	-2.43	-0.003	-0.06	-0.1642	-4.70
X23 (X2 X3 interactions)	0.0164	3.44	0.002	0.17	0.0944	1.84	-0.3059	-8.83

3. Experienced drivers are not as willing to accept advice as less experienced drivers, and they also make their decisions faster.
4. A "freeway bias" exists with subjects more willing to accept freeway advice.
5. Although males are more willing to accept advice, they are also less likely to purchase an information system.
6. Whereas less experienced drivers are more likely to follow advice, they are also less likely to purchase an information system.

MODELING SEQUENTIAL ROUTE CHOICE BEHAVIOR

The ultimate goal of this research effort is to develop a realistic model of route choice behavior under the influence of ATIS, which incorporates the effects of drivers' learning abilities. Two modeling approaches are under investigation for the development of a route choice behavioral model as part of this research effort. The first approach, which is described here, is the use of a conventional logit model formulation. The second approach is the use of a neural network model and is described elsewhere (*7*). The use of the logit model and the random utility theory assumes that an individual's choice between two or more alternatives is based on the utility gain experienced by the individual for a particular choice. The reliable estimation of an individual's perceived utility, then, is of primary importance in estimating the overall model. The individual's perceived utility of an alternative is used in lieu of the actual utility because, although the actual utility may be greater or less than the perceived utility, it is an individual's perception of reality that ultimately drives behavioral responses.

It is reasonable to assume that an individual's perceived utility for a specific alternative is a function of the perceived attributes of the alternative, an individual's characteristics (personal biases or preferences), the information available on the alternative, and the perception of the accuracy of such information. There may also be an effect on the perceived utility of an alternative because of a repetitive choice effect. Simply stated, the more times one chooses an alternative the greater the perceived utility becomes for that alternative because of some habitual nature of the individual. This general framework forms the basis for the formulation of specific alternative utility functions within this analysis.

When analyzing sequential choices, the utility functions for each alternative must be updated to reflect the individual's learning processes. Thus, the perceived utility for a specific alternative for a given trial is dependent on the perceived outcome of previous trials or experiences. Each sequential choice results in an experience, which in turn influences the next choice. Just how much this past experience affects the current choice and how rapidly individuals modify their behavior on the basis of their experiences will give an indication of the learning abilities of the individual. Various information updating strategies exist, and finding the most appropriate formulation to apply to drivers' route choice behavior may require a certain amount of trial and error, if an appropriate formulation can be found at all. If the learning and adaptive abilities of drivers vary greatly, then it may be impossible to specify an appropriate updating function that applies to a majority of drivers. The information updating strategy used in this analysis is as follows:

$$x_{ij}(k) = \tau x_{ij}(k - 1) + (1 - \tau)u_{ij}(k - 1) \qquad (2)$$

where $x_{ij}(k)$ is the perceived value of attribute x by individual i for alternative j for trial k, and, likewise, $x_{ij}(k - 1)$ is the perceived value for the previous trial $k - 1$.

Thus defined, $x_{ij}(k)$ becomes an endogenous variable. At this early stage of the modeling effort the issue of endogeneity has been ignored. The variable $u_{ij}(k - 1)$ is the actual value of the attribute as experienced by individual i on the previous trial $k - 1$. The coefficient τ is an experience importance factor whose value gives an indication of the relative importance of an individual's previous experiences in updating perception or expectation on the current trial. Such linear combinations are often used in network assignment.

For an individual who has not performed any trials, no previous experiences exist; therefore a perception of the attribute cannot be developed from previous experiences. Individuals may, however, have preconceived perceptions of certain attributes, in which case initial conditions must be established for the individual attributes. In the route choice simulation used to collect data for this analysis, subjects were given a significant amount of preliminary information about the simulation such that they could develop some initial perceptions of simulation attributes; therefore, initial conditions were established for individuals' perceptions of some attributes.

BINARY LOGIT MODEL

The sequential route choice processes were modeled using the binary logit formulation (*8*). The random utility function is the perceived utility of person i, for alternative route j, on the k^{th} day and is defined as follows:

$$V_{ij}(k) = \beta_0 + \sum_{l=1,5}\beta_l X_{ijl}(k) + \varepsilon_{ij}(k) \qquad (3)$$

where

$X_{ij1}(k)$ = a dummy variable indicating which route is the advised route. It is 1 when the advised route is alternative j for day k and is 0 otherwise;
$X_{ij2}(k)$ = the perceived delay on alternative j for individual i for day k;
$X_{ij3}(k)$ = the perceived accuracy level of the information provided for the advised route;
$X_{ij4}(k)$ = a dummy variable, 1 if the subject is male for the advised route; and
$X_{ij5}(k)$ = a dummy variable, 1 if the subject is an inexperienced driver for the advised route.

The utility function for alternative j then is simply a linear combination of an alternative-specific coefficient β_0, the above variables, and an independent extreme-value distributed error term $\varepsilon_{ij}(k)$. The first variable represents the increase in utility of the advised route over the remaining alternative. This formulation assumes that, by advising a subject to take a specific route, the perceived utility for that route increases, thus increasing the probability of choosing that route. On the basis of the ANOVA results that showed an individual agreement with advice of about 72 percent, this variable should contribute significantly to the utility function and have a positive coefficient.

The second variable is an experience variable that represents an individual's perception of the delay to be experienced on either the side road or the freeway. This perception of delay must be updated for each sequential trial to incorporate the learning process on the basis of previous experiences. This variable is updated using the previously described updating function. At the beginning of the route choice simulation, subjects are allowed to view the fastest possible travel times on the freeway and side road, and likewise the slowest possible times. This in effect creates an initial perception of the delay to be experienced on the individual routes. For this analysis, the average of the minimum and maximum possible delay, as displayed to the subject, was used as the initial perceived delay for the two alternative routes.

The third variable is another experience variable representing an individual's perception of the accuracy level of the information being provided. Subjects are told at the start of the simulation that their "traffic watch" device will not always be accurate but are not given any indication of the overall accuracy of the device. It is reasonable to assume that in the absence of any other information, subjects will assume that the information being provided is correct until, through an accumulation of their experiences, they develop a perception of the accuracy of the information system. For this analysis, subjects are initially assumed to perceive the information as being 100 percent accurate, and then their perception is updated using the updating function to account for the effects of their experiences.

The fourth and fifth variables are personal attribute variables that the ANOVA indicates have strong effects on the individual's acceptance of advice. These variables result in increasing the utility of the advised route for male subjects and inexperienced drivers, thus increasing the probability that the subject will accept the advice. It was shown in the ANOVA that these two characteristics resulted in a higher average acceptance rate of advice for subjects with these characteristics.

An alternative specific coefficient was included for the freeway alternative. The ANOVA results indicated a preference or bias toward the freeway indicating that, all else being equal, the perceived utility for the freeway alternative should be greater than that for the side road. It is expected then that this freeway coefficient should be positive.

The model specified above was estimated over a range of experience importance factors with $0 \leq \tau \leq 1$ and with 0.2 increments. The data set for this model included 1,376 individual choices made by 43 subjects (23 from Experiment 1, Condition 2, and 20 from Experiment 3, Condition 7) all of which were subjected to the same experimental conditions. The model estimation technique uses the maximum log-likelihood method. The estimated model coefficients and log-likelihood values are presented in Table 3.

RESULTS

Of the six models estimated and presented in Table 3, the model with the greatest log-likelihood value is the model for which τ was set equal to 0.8. Of the models that incorporate some amount of utility on the basis of experience ($\tau < 1.0$), this model is also the only model in which all the coefficients have the appropriate sign. For the first variable, which represents the effects of route information, the positive value of 0.409 indicates that there is an increase in utility for the advised route and thus an increase in the probability that this route will be selected. The t-statistic for this variable indicates that the coefficient is not individually significantly different from 0 ($t < 1.96$ @ $\alpha = 0.05$), indicating caution in the interpretation of this variable. The second variable, which is

TABLE 3 Logit Model Coefficients

	$\tau = 0$		$\tau = 0.2$		$\tau = 0.4$		$\tau = 0.6$		$\tau = 0.8$		$\tau = 1.0$	
	β	t	β	t	β	t	β	t	β	t	β	t
$X_{ij1}(k)$ Advised route dummy entering alternative 1 and 2	1.199	7.15	1.136	6.06	0.998	4.61	0.741	2.75	0.409	1.01	1.135	9.90
$X_{ij2}(k)$ perceived delay entering alternative 1 and 2	0.059	1.67	0.065	1.49	0.065	1.18	0.047	0.62	-0.039	-0.33	-	
$X_{ij3}(k)$ perceived accuracy of advice entering alternative 1 and 2	-0.094	-0.58	-0.009	-0.05	0.174	0.71	0.511	1.59	0.922	1.86	-	
$X_{ij4}(k)$ male gender dummy entering alternative 1 and 2	0.294	2.00	0.293	2.00	0.292	1.99	0.294	2.00	0.296	2.02	0.290	1.98
$X_{ij5}(k)$ inexperienced driver dummy entering alternative 1 and 2	0.331	2.41	0.328	2.39	0.326	2.38	0.327	2.38	0.334	2.43	0.325	2.37
β_0 freeway alternative specific coefficient	0.514	7.21	0.514	7.21	0.514	7.20	0.513	7.17	0.503	6.87	0.510	7.17
L(β) log-likelihood at convergence	-670.03		-670.40		-670.46		-669.91		-669.79		-671.52	
L(0) log-likelihood with all coefficients equal to zero	-953.78											
L(C) log-likelihood with all coefficients equal to zero except β_0	-936.11											

an updated perception of the delay on the alternative routes, has a coefficient value of -0.039, indicating that as the perceived delay on a route increases, the utility of that route decreases. Again, the *t*-statistic indicates that this coefficient is not individually significantly different from 0. The third variable, which is an updated perception of the accuracy of the information, has a coefficient value of 0.922, indicating that as the perceived accuracy of the system increases, the utility of the advised route increases. The *t*-statistic for this variable as well indicates that this coefficient is not individually significantly different from 0. The last two coefficients, which are indicators of the subject's sex and driving experience level, both have positive values and have *t*-statistics that indicate individual significance, as was expected on the basis of the ANOVA results. The freeway alternative-specific coefficient has the expected sign and is individually significant, again reiterating the "freeway bias" of subjects. This coefficient and its associated *t*-statistic remained relatively constant across all estimated models.

The overall fit of this model is not significantly different from the fit of any of the other models, as indicated by the relatively small variation in the maximum log-likelihood values. This brings into question the collective significance of coefficients β_2 and β_3 and the relative importance of the effects of previous experiences on current choices. When $\tau = 1.0$ in the updating function, there is no effect of previous experiences included in an individual's perception of information and route attributes for the current choice. The model for $\tau = 1.0$ was estimated by dropping these two variables from the analysis. The model estimated for $\tau = 1.0$ can then be used to test the collective significance of β_2 and β_3. If the $L(\beta)$ for this last model is defined as the log-likelihood for which coefficients β_2 and β_3 are constrained to 0 ($\beta_2 = \beta_3 = 0$) and is identified as $L(\beta_c)$ then the value $-2[L(\beta_c) - L(\beta)]$ has a chi-square distribution with two degrees of freedom and can be used to test the collective null hypothesis. From the values in Table 3, this chi-square value can be calculated as 3.46, which does not provide evidence to reject the null hypothesis ($X^2 > 5.99$, df $= 2$, $\alpha = 0.05$).

These results indicate two possibilities. The first is that drivers' perceptions of attributes, based on previous experiences, have little effect on route choice behavior under the influence of ATIS or it may be simply that the updating function used in this analysis is flawed and does not accurately describe the updating processes of drivers. The statistical tests of the coefficients for the updated variables, indicating collective insignificance, support both of the above hypotheses. The model with $\tau = 1.0$ includes only the system advice, personal attributes, and the alternative-specific coefficient, yet still predicts the route choice behavior fairly well with 79.2 percent of the 1,376 choices accurately predicted. The prediction rates for this model are presented in Table 4.

These results indicate that an accurate model of route choice behavior, exclusive of learned attributes, may be possible. The model prediction rate of 79.2 percent is approximately the same as the average acceptance rate of advice and the accuracy of the advice. The model then may be simply predicting that the route chosen is the route advised, and this is evidenced by the strong significance of the advice variable ($t = 9.9$) and the relative size of its coefficient. Counter to this argument is the fact that the model prediction rate is much better for the side road than for the freeway and that the prediction rate for the side road is significantly higher than the acceptance rate for the side road, which indicates that the model is not only predicting what is advised. Although excluding experience effects seem to be a gross simplification of the route choice behavioral process, it may be an accurate simplification. If it can be determined that drivers will follow route advice consistently at a rate equivalent to the accuracy of the advice being provided, this simple model of route choice may be adequate for use in a traffic assignment model. Only one possible updating function was used in this analysis, which may be why there were no significant effects as a result of updated perceptions. Continued modeling efforts will be undertaken using various updating schemes to determine the significance of drivers' learning experiences.

SUMMARY AND CONCLUSIONS

Previous research by the authors (*1*) has shown that a basic understanding of drivers' route choice behavior is necessary to develop predictive models of drivers' en route diversion choice. To study the basic underlying factors that contribute to diversion behavior, an interactive computer simulation experiment was developed in an attempt to capture drivers' sequential learning processes. Analysis of these experimental data resulted in the discovery of some interesting relationships. The ANOVA findings provide evidence that males will accept route advice more often than females over a range of accuracies and that inexperienced drivers will follow route advice more often than experienced drivers. In contrast to these findings, when asked about potential use of such an ATIS device, females were more likely to purchase an information device when accuracies were at 60 and 75 percent, but at 90 percent accuracies, males were more likely to purchase a device. These findings indicate that although males accept advice more readily at all accuracy levels, they are not as willing to purchase such a device unless the system is very accurate. A similar finding related to driving experience also

TABLE 4 Model Prediction Rate

		Predicted Choices		Total	% Correctly
		Freeway	Side Road	Number	Predicted
Actual Choices	Freeway	600	198	798	75.2%
	Side Road	88	490	578	84.8%

exists. Although inexperienced drivers were more likely to follow the advice being given, they also reported being less likely to purchase such an information device. This may be the result of less frequent drivers feeling that the savings gained from such a device would not outweigh the costs because of their limited driving. Conversely, more experienced and more frequent drivers perceive a net gain and respond as more likely to purchase a device although they do not follow the advice as often. The ANOVA also revealed that drivers will follow advice to take the freeway more readily than advice to take the side road and that they are quicker to respond to freeway advice, indicating that a "route bias" exists.

Analysis of the route choice decision times of drivers found that there was a very rapid drop in the decision times over the first 8 of 32 trials and that the times remained relatively constant over the remaining 24 trials. This finding and the fact that average acceptance rates of advice approximated the accuracy of the system indicate that drivers could sense and adapt quickly to the level of accuracy being provided by the system. Average decision times were the greatest for information provided at 75 percent accuracy. This indicates that subjects were more readily able to identify the level of accuracy for low levels as well as high levels but took a greater amount of time to discern the moderate level of accuracy.

The efforts to develop a model of route choice behavior that incorporates the learning processes of drivers had mixed results. A model was developed that included drivers' updated perceptions of route delay and information accuracy, but the model was not significantly different from a model that excluded these perceived attributes. The model includes the advised route as a variable. Because subjects followed the advice so readily, the model may simply be predicting that subjects will select the advised route, therefore predicting about 79 percent correct, which is equivalent to the average acceptance rate of advice. More analysis is required using different updating schemes before conclusive results can be made about the effects of experiences on sequential trials. Future research efforts will include attempts to formulate more realistic information updating schemes and to extend the research and modeling effort to a more realistic traffic network environment.

ACKNOWLEDGMENTS

The authors wish to thank the California Department of Transportation and the Partners for Advanced Transit and Highways for funding this research.

REFERENCES

1. M. A. Abdel-Aty, K. M. Vaughn, R. Kitamura, and P. P. Jovanis. *Impact of ATIS on Drivers' Travel Decisions: A Literature Review*. Research Report. UCD-ITS-RR-92-7. Institute of Transportation Studies, University of California, Davis, 1992.
2. P. Bonsall and M. Joint. *Driver Compliance with Route Guidance Advice: The Evidence and its Implications*. Society of Automotive Engineers, Warrendale, Pa., Oct. 1991, pp. 47–59.
3. P. Bonsall. The Influence of Route Guidance Advice on Route Choice in Urban Networks. Special Issue. *Proc., Japanese Society of Civil Engineers*, Feb. 1991.
4. P. Bonsall and T. Parry. A Computer Simulation Game to Determine Drivers' Reactions to Route Guidance Advice. *Proc., 18th PTRC Summer Annual Meeting*, London, 1990.
5. J. Neter, W. Wasserman, and M. H. Kutner. *Applied Linear Statistical Models*. IRWIN, Homewood, Ill., 1990.
6. K. M. Vaughn, M. A. Abdel-Aty, R. Kitamura, and P. P. Jovanis. *Experimental Analysis of Sequential Route Choice Behavior Under ATIS*. Research Report UCD-ITS-RR-92-16. Institute of Transportation Studies, University of California, Davis, 1992.
7. H. Yang, R. Kitamura, P. P. Jovanis, K. M. Vaughn, and M. A. Abdel-Aty. Exploration of Route Choice Behavior with Advanced Traveler Information Using Neural Network Concepts. In *Transportation*, Kluwer Academic Publishers, 1993.
8. M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis Theory and Application to Travel Demand*. The MIT Press, Cambridge, Mass., 1985.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Network Performance Under System Optimal and User Equilibrium Dynamic Assignments: Implications for Advanced Traveler Information Systems

HANI S. MAHMASSANI AND SRINIVAS PEETA

A comparative assessment was undertaken of network cost and performance under time-dependent system optimal (SO) and user equilibrium (UE) assignment patterns, with particular reference to the effectiveness of advanced traveler information systems (ATIS). Both SO and UE solutions were found using a new simulation-based algorithm for the time-dependent assignment problem. Experiments were conducted using a test network with signal-controlled junctions under progressively increasing network loading intensities. A diagnosis of system performance for various intensities of loading was effected using network-level traffic descriptors for both SO and UE assignments. The results affirm the validity of a meaningful demarcation between SO and UE assignments in urban traffic networks and provide useful insights for macroscopic network-level relations among traffic descriptors. These results suggest that ATIS information supply strategies based on SO route guidance could considerably outperform descriptive noncooperative information strategies, especially at moderate to high congestion levels in the network. The results also illustrate the time-dependent nature of the gains achieved by an SO assignment vis-à-vis a UE assignment in a congested traffic network.

Approaches incorporating advances in communication technologies, information processing systems, electronics, and automation, broadly labeled as intelligent vehicle highway systems (IVHS), continue to generate considerable interest for their potential to alleviate urban and suburban congestion of traffic systems. Advanced traveler information systems (ATIS) provide travelers with real-time information on existing traffic conditions or instructions on route selection from their current location to their destinations, or both. Successful implementation of ATIS, especially at high market penetration levels, involves the dynamic assignment of vehicles to "optimal" paths to reduce overall system user costs. Recently Mahmassani and Peeta (*1*) proposed a heuristic algorithm to solve the system optimal (SO) dynamic traffic assignment problem for the ATIS context, in which a central controller with known or predicted time-dependent origin-destination (O-D) trip desires over the horizon of interest solves for paths to prescribe to users to attain some systemwide objectives. A comprehensive review and discussion of dynamic assignment and traffic simulation models for ATIS-advanced traveler management system (ATMS) applications are given by Mahmassani et al. (*2*).

Department of Civil Engineering, The University of Texas at Austin, Austin, Tex. 78712-1076.

In this paper, the performance of a traffic network employing this solution methodology is analyzed for both SO and user equilibrium (UE) time-dependent assignments. As in the static case, SO and UE dynamic assignments involve similar algorithmic steps, differing primarily in the specification of path travel costs that form the basis of the corresponding assignments. SO dynamic assignment is accomplished using time-dependent marginal travel times [see Ghali and Smith (*3*)], whereas a UE assignment is attained using the time-dependent average travel times. The system performance under the above assignment schemes was analyzed for various intensities of network loading covering the spectrum of network states from uncongested networks to very highly congested networks. In addition, the numerical experiments illustrate the extent of the differences between SO and UE time-dependent assignments in terms of total system cost at varying levels of network congestion. This question is of fundamental importance to ATIS operations, with regard to the relative benefits of normative versus descriptive information supply strategies.

An SO assignment does not generally represent an equilibrium flow pattern because some users may be able to obtain individual advantages simply by changing routes, though they may impose a greater marginal cost on other users in the system in the process. Its significance to the ATIS context lies in providing a benchmark against which other assignments or flow patterns can be gauged, thereby yielding an upper bound on the benefits attainable with real-time traffic information. A Wardrop UE holds when users cannot improve their individual costs by unilateral route switching. There is no empirical evidence that UE conditions hold in real networks, although the UE solution is considered a reasonable and useful construct for the evaluation of long-term capacity improvements. Under real-time descriptive ATIS information on network conditions, a time-dependent UE pattern could be viewed as the result of the long-term evolution of the system, as users somehow learn and adjust under the supplied information. However, it is not at all clear that such convergence would be attained under inherently dynamic conditions (exacerbated by supplying information to users). Thus it is not known what the UE solution may represent from the standpoint of ATIS operation and evaluation. Actual user behavior and system performance under real-time descriptive information may be better or worse than the corresponding time-dependent UE solution in terms of the overall system cost. Nevertheless, a time-dependent UE pattern may be considered as a useful proxy for a favorable scenario of long-term network performance under real-time descriptive information.

It is known from static network equilibrium theory that SO and UE lead to identical solutions only for situations in which the shortest paths taken by users simultaneously are the best paths from a system viewpoint. Such situations are observed when networks are relatively uncongested so that link operating speeds are unaffected by the flows on the links (limited vehicle interactions). At the other extreme, under highly congested conditions, system performance is not likely to be markedly different under the two assignment schemes because the opportunities for SO to sufficiently ameliorate the traffic situation probably would be limited.

For network conditions between the two extremes, the extent of the differences between SO and UE solutions, particularly in terms of overall system cost, is not known. This is very important for ATIS because if the two solutions are not perceptibly different, coordinated cooperative SO route guidance imposed by a central controller may not be necessary, and descriptive information that is less complicated and simpler to disseminate to noncooperating drivers may be sufficient. The similarity of the two solutions would have important implications for the focus that ATIS information supply strategies should take, with more attention directed to ways of guiding the system toward UE convergence and away from wide fluctuations. However, if SO indeed holds promise for meaningful gains over UE, then normative route guidance or strategies to induce the system near its SO should be pursued. It is also desirable to ascertain network and traffic conditions under which differences between SO and UE are meaningful.

In this paper, overall user cost and network performance under time-dependent SO and UE assignment patterns are examined in a series of numerical experiments performed on a test network under various loading levels. The system performance is gauged using average network-level traffic flow descriptors, in addition to the standard parameters such as average travel time. The time-dependent nature of the problem further complicates the already intricate problem of characterizing traffic flow performance at the network level that was previously addressed only under steady-state conditions, as discussed hereafter.

NETWORK TRAFFIC FLOW THEORY

Mahmassani et al. (*4,5*) generalized the definitions of speed, flow, and concentration to the network level and examined their interrelation in their model of network traffic performance. These concepts are extended to the dynamic case in the current analysis to characterize the vastly varying network traffic conditions (especially for medium to high network loading levels) during the peak period. Average network speed V (kilometers per hour) is obtained as the ratio of total vehicle kilometers to total vehicle hours in the network over the duration of interest. The average network concentration K (vehicles per lane kilometers), for the duration of interest, is the time average of the number of vehicles per unit lane length in the system. However, the concentration varies dramatically with time in dynamic traffic networks. Hence, the time-dependent network concentration is examined by taking 5-min averages of the number of vehicles per unit lane length in the system. An overall measure of network concentration K over the duration of the period of interest is obtained by taking the arithmetic average of the 5-min averages. Similarly, time-dependent network flow, interpreted as the average number of vehicles per unit time that pass through a random point along the network, is examined by taking 5-min averages; an overall measure of network flow Q over the peak period is obtained by taking the simple average of $(\Sigma l_i q_i)/(\Sigma l_i)$, where q_i and l_i, respectively, denote the 5-min average flow and the length of Link i, and the summations are taken over all network links.

Two fundamental relationships between these three network traffic flow variables are investigated in this study. The first relates average network speed (V) and average network concentration (K). For arterials or single roadways, a qualitative trend of decreasing speed with increasing concentration is well established. The same general trend was observed to hold at the network level in the simulation experiments of Mahmassani et al. (*5*), although the complexity of network interactions preclude the analytic derivation of such a relation directly from the link-level relations. The second relationship analyzed is the basic identity $Q = KV$. Formally established for single roadways, it was shown to also hold at the network level in the previously mentioned steady-state experiments (*5*). These experiments were performed keeping the network concentration level constant for the duration of interest by treating the network as a closed system. The NETSIM package was used for the study, and vehicular behavior was governed by the comprehensive microscopic rules embedded in NETSIM. The present study replicates the network traffic conditions of a rush hour traffic situation. It uses the DYNASMART (DYnamic Network Assignment Simulation Model for Advanced Road Telematics) simulation-assignment model developed at The University of Texas at Austin for ATIS-ATMS applications. The $Q = KV$ identity is expected to hold only approximately for time-varying network traffic flow.

SOLUTION METHODOLOGY

Problem Statement

Consider a traffic network represented by a directed graph $G(N, A)$ where N is the set of nodes and A the set of directed arcs. A node can represent a trip origin, a destination, or a junction of physical links. Consider a network with multiple origins and destinations. The time experienced by a vehicle to traverse a given link depends on the interactions taking place among vehicles in the traffic stream along this arc. The analysis period of interest, taken here as the peak period, is discretized into small equal intervals $t = 1, \ldots, T$. Given a set of time-dependent O-D vehicle trip desires for the entire duration of the peak period, expressed as the number of vehicle trips r_{ijt} leaving Node i for Node j in time slice t, $\forall\ i, j \in N$ and $t = 1, \ldots, T$, determine a time-dependent assignment of vehicles to network paths and corresponding arcs. In other words, find the number of vehicles r^k_{ijt} that follow path

$k = 1, \ldots, K_{ij}$ between i and j at time t, $\forall\, i, j \in N$ and $t = 1, \ldots, T$, as well as the associated numbers of vehicles on each arc $l \in A$ over time. As explained in the previous section, two such assignments are computed: (a) one that satisfies UE conditions that no user can improve actual (experienced) trip time by unilaterally changing routes and (b) one that minimizes total travel time (for all users) in the system over the peak period. The interpretation of these two solutions from the standpoint of ATIS effectiveness was discussed in the previous section.

Simulation Assignment Solution Procedure

This section describes briefly the algorithm used to solve for SO and UE assignments. A detailed description of the solution procedure by Mahmassani and Peeta (*1*) consists of a heuristic iterative procedure in which a special-purpose traffic simulation model is used to represent the traffic interactions in the network and thereby evaluate the performance of the system under a given assignment. As indicated earlier, the algorithmic steps for UE assignment are virtually identical to those for the SO solution except for the specification of the appropriate arc costs and the resulting path processing component of the methodology. The algorithm is first summarized for the SO case, followed by a brief description of the modification for the UE problem.

The use of a traffic simulation model to evaluate the SO objective function and model system performance circumvents the principal difficulties that have precluded solutions to realistic formulation of the problem by obviating the need for link performance functions and link exit functions and implicitly ensuring that the first-in, first-out property holds on traffic facilities and that no unintended holding back of traffic takes place at nodes [see Mahmassani et al. (*2*) for a discussion of issues arising in dynamic traffic assignment]. The algorithm uses the DYNASMART simulation assignment model. DYNASMART has the capability to simulate the movement of individual vehicles through the network, with path selection decisions possible at every node or decision point along the way to the destination, as supplied by the user decision rules reflecting driver behavior in response to real-time information. In this work, vehicular paths are preassigned exogenously to DYNASMART, as determined by the steps of the SO or UE solution algorithms. Thus DYNASMART is used primarily as a simulator to replicate the dynamics of traffic phenomena in response to a given assignment of vehicles to paths. A detailed description of the various capabilities of DYNASMART is provided by Mahmassani et al. (*6*).

The simulation results provide the basis for a direction-finding mechanism in the search process embodied in the solution algorithm for this nonlinear problem. The experienced vehicular trip times from the current simulation are used to obtain a descent direction for the next iteration. The time-dependent shortest travel time paths and least marginal travel time paths are obtained using the time-dependent algorithms described by Ziliaskopoulos and Mahmassani (*7*). An elegant aspect of the solution methodology is that it avoids complete path enumeration between O-D pairs.

Figure 1 depicts the solution algorithm for the SO dynamic traffic assignment problem. A brief summary of the approach is as follows:

1. Set the iteration counter $I = 0$. Obtain the time-dependent historical paths (paths obtained from the data base) for each assignment time step over the entire duration for which the assignment is sought.
2. Assign the O-D desires (which are known a priori for the entire peak period) for the entire duration to the given paths and simulate the traffic pattern that results from the assignment using DYNASMART.
3. Compute the marginal travel times on links using time-dependent experienced or estimated link travel times and the number of vehicles on links obtained as post-simulation data (from Step 2).
4. Using a special-purpose, time-dependent, least-cost path algorithm, compute the least marginal time paths for each O-D pair for each assignment time step on the basis of the marginal travel times obtained in Step 3.
5. Perform an all-or-nothing assignment of O-D desires to the least marginal time paths computed in the previous step. The result is a set of auxiliary path vehicle numbers for each O-D pair for each assignment time step $t = 1, \ldots, T$.
6. Update paths and the number of users assigned to those paths. Paths are updated by checking whether the path identified in Step 4 already exists (i.e., has carried vehicles in at least one prior iteration) for that O-D pair and including it if it does not. The update of the number of vehicles (assignment of vehicles to the various paths currently defined between the O-D pair after the path update) is performed using the method of successive averages (MSA), which takes a convex combination of the current path and corresponding auxiliary path numbers of vehicles for each O-D pair and each time step. A detailed description of MSA is provided by Sheffi and Powell (*8*). Note that other convex combination schemes could be used equally.
7. Check for convergence using an ε-convergence criterion.
8. If the convergence criterion is satisfied, stop the program. Otherwise, update the iteration counter $I = I + 1$ and go to Step 2 with the updated data on paths and the number of vehicles assigned to each of those paths.

The complexity of the interactions captured by the simulator when evaluating the objective function generally precludes the kind of well-behaved properties required to guarantee convergence of the algorithm in all cases. However, such convergence was achieved in all the experiments reported in this paper and in many other test networks solved to date. Also, path marginals are not necessarily global because they are based on link level marginal travel times. Efforts were made to attain a global optimum where local solutions were suspected.

Modification To Obtain User Equilibrium Solution

As previously discussed, the solution to the time-dependent UE problem is obtained by assigning vehicles to the shortest average travel time paths instead of the least marginal paths in the direction-finding step (Step 5). In other words, use the

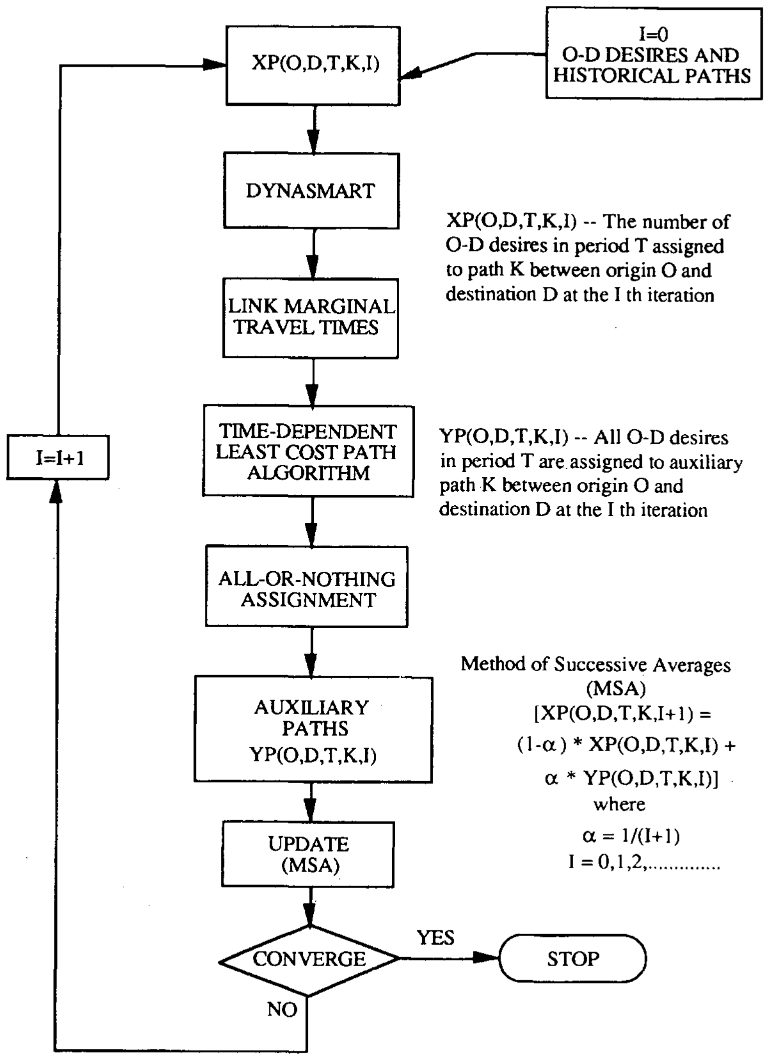


FIGURE 1 Solution algorithm for the SO dynamic assignment problem.

(time-dependent) average travel times on links instead of the marginal travel times in the shortest-path calculations. In the above solution procedure, this simplifies Step 3 and modifies Step 4 as indicated.

EXPERIMENTAL DESIGN AND SETUP

Network Configuration and Traffic Characteristics

The test network used in this study, having 50 nodes and 163 links, consists of a freeway with a street network on both sides, as shown in Figure 2. Nodes within the freeway section are neither origin nor destination nodes. A total of 38 origin nodes and 38 destination nodes are obtained by excluding freeway nodes (Nodes 1 through 37 and 44). Freeway nodes are connected to the street network through entrance and exit ramps. Unless otherwise indicated in Figure 2, all arcs shown are two directional. All links are 0.83 km (0.5 mi) long and have two lanes in each direction, except for the entrance and exit ramps, which are directed arcs with a single lane. The freeway links have a mean free speed of 91.67 km/hr (55 mph), and the other links have a mean free speed of 50 km/hr (30 mph). In terms of traffic signal characteristics, 25 intersections have pretimed signal controls, 8 have actuated signal controls, and the remaining 17 nodes have no signal control.

Experimental Setup

The comparative assessment of system performance for SO and UE assignments is conducted under various network loading levels, which generate various levels of network congestion. The network loading factor is defined as the ratio of the total number of vehicles generated in the network during the assignment period to a given reference number (19,403 vehicles over a 35-min period in the experiments). Table 1 shows the various loading factors considered in this study and the corresponding number of vehicles generated on the test network during the duration of interest (35 min in all cases). In addition, it shows the corresponding number of "tagged" vehicles (vehicles generated for the 30-min duration after the

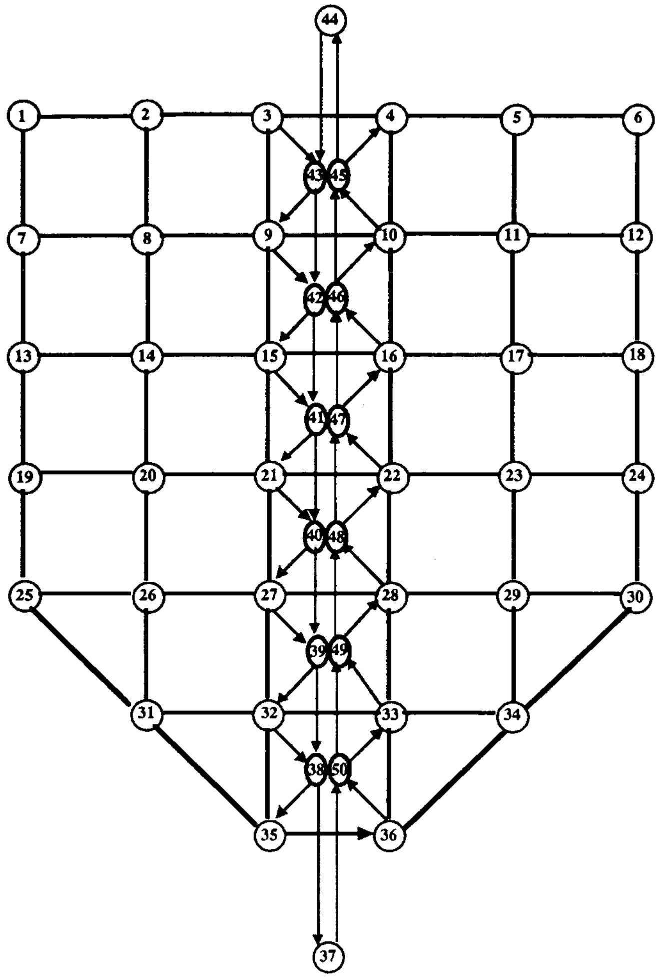


FIGURE 2 Network structure.

5-min start-up time) for which relevant performance statistics are accumulated. The loading factors range from 0.6 (very low congestion with 11,616 vehicles) to 2.4 (extremely high congestion with 46,674 vehicles). Under each loading level, the UE and SO solutions are obtained, and the resulting time-dependent link flow patterns are obtained from DYNASMART. The shape of the loading curve for the various network loading levels emulates real-world network loading for the peak period, with an initially increasing generation rate until a peak is reached, followed by a decreasing vehicle generation rate.

In the present study, a start-up time of 5 min is provided in DYNASMART for the network to be reasonably occupied, followed by a 30-min peak period generation of traffic (for which performance statistics are accumulated). Another aspect of the experimental setup that critically influences the system performance is the spatial distribution of the O-D demand pattern. The vehicles generated are about evenly distributed spatially in terms of both their origins and destinations, except for Nodes 37 and 44, which generate or attract only about 25 percent of the number of vehicles originating or destined to a typical O-D node (i.e., Nodes 1 through 36).

RESULTS

The results from the various experiments are viewed from two principal perspectives. First, they form the basis for comparison of system performance, particularly user costs under UE and SO assignment schemes, thereby addressing the questions relevant to ATIS information strategies described in the introductory section of the paper. Second, they are used to investigate network-level traffic flow characteristics and relations using network-wide traffic descriptors. This investigation is conducted primarily for the SO flow pattern. An additional element of the study is the time-dependent analysis of the travel time gains of SO over UE, also of significance to ATIS operation.

The results provide several key insights from both of the above perspectives. They manifest a clear qualitative and quantitative distinction in the solution provided by the SO assignment scheme as opposed to the time-dependent UE assignment procedure to route vehicles in a general traffic network. The results also reveal important and robust macroscopic relationships among network-level traffic variables that parallel those for single roadways.

System Performance for SO and UE Assignments

Table 1 reports summary statistics on the system performance for the SO and UE assignments for the various loading factors. As expected, at low levels of network loading, when the network is relatively uncongested, the average travel times of vehicles in the network are relatively close across the various loading levels. As the load is increased, the effects of congestion become more prominent and the average travel times in the network increase at an increasing rate with the loading factor. At very high loading levels, the marginal effect of additional demand on system performance is very high. The results also indicate that there is only limited variation in the average distance traveled by vehicles under the various network loading levels, implying that greater congestion and not longer travel routes is the primary cause of the higher system trip times (the objective function seeks to minimize total system travel time only). Nevertheless, the average travel distance does increase with the loading level, reflecting an increasing percentage (although small in magnitude) of drivers assigned to longer travel routes. The average travel distances under UE for various network loading levels are smaller than the corresponding distances for SO, indicating a smaller percentage of long travel routes under UE. This may be explained by some users being assigned to longer routes to reduce congestion elsewhere to reduce systemwide travel times.

Figure 3 shows comparatively the average trip times under various network loads for UE and SO assignments. As discussed above, both curves illustrate the increasing marginal effects of additional demand on system trip times. Of more relevance to the central question addressed in this paper, Figure 3 highlights the difference in the quality of the solutions provided by the two assignment rules for time-dependent network flows. This is further illustrated in Figure 4, which depicts the percentage improvement in average travel time of SO over UE (as a fraction of the UE travel time) for the various average network concentrations corresponding to the

TABLE 1 Summary Statistics for SO and UE Assignments

Loading Factor	No. of Generated Vehicles	No. of Tagged Vehicles	Average Trip Time (minutes)	Total Trip Time (hours)	Average Trip Distance (km)	Total Trip Distance (km)	Average Speed (kmph)
System Optimal							
0.60	11616	10585	3.85	679.5	3.03	32096.2	47.23
0.80	15509	14098	3.90	916.0	3.02	42411.6	46.30
1.00	19403	17621	4.03	1183.1	3.03	53391.6	45.22
1.20	23305	21145	4.40	1549.5	3.07	64728.8	41.77
1.40	27196	24697	4.86	1999.1	3.08	76207.9	38.12
1.60	31090	28205	6.04	2837.1	3.20	90222.1	31.80
1.80	34978	31726	7.65	4042.9	3.28	103997.5	25.72
2.00	38871	35258	10.46	6149.5	3.32	117013.3	19.03
2.10	40818	37014	13.08	8071.9	3.35	123996.7	15.37
2.20	42769	38784	16.57	10710.9	3.32	128811.7	12.03
2.40	46674	42322	24.95	17601.7	3.55	149978.3	8.52
User Equilibrium							
0.60	11616	10585	3.86	681.5	3.00	31839.6	46.72
0.80	15509	14098	3.92	920.8	2.97	41898.3	45.50
1.00	19403	17621	4.15	1219.5	2.98	52656.2	43.18
1.20	23305	21145	4.60	1622.5	3.02	63731.2	39.28
1.40	27196	24697	5.43	2236.5	3.00	74289.6	33.22
1.60	31090	28205	6.79	3192.2	3.08	87165.4	27.30
1.80	34978	31726	9.00	4762.9	3.13	99513.3	20.88
2.00	38871	35258	12.91	7587.7	3.27	115249.2	15.18
2.10	40818	37014	14.94	9215.7	3.22	119132.5	12.93
2.20	42769	38784	18.55	11993.6	3.32	128605.0	10.72

NOTE: 1 km = 0.6 mile

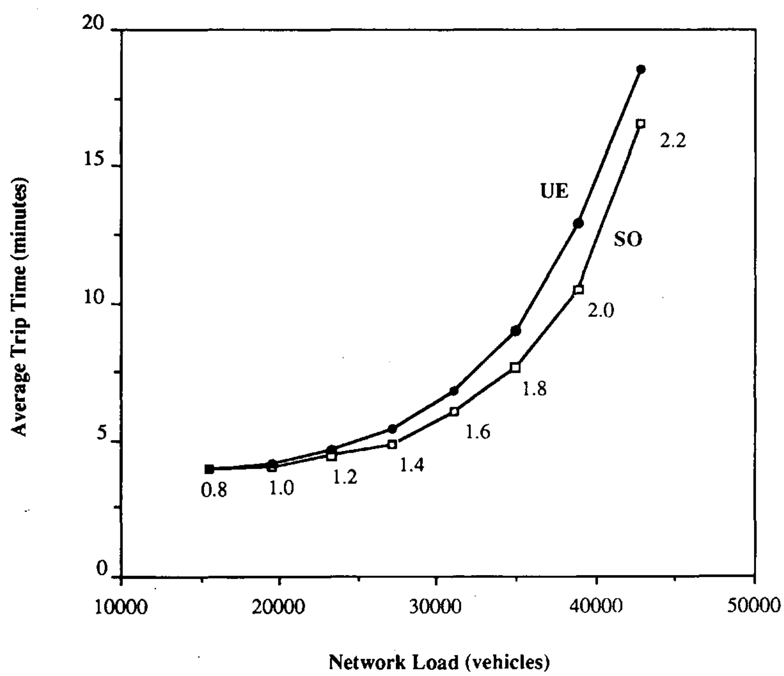


FIGURE 3 Comparison of average trip times (minutes) of SO and UE assignments for various levels of network loading. The number by each plotted point is the corresponding loading factor.

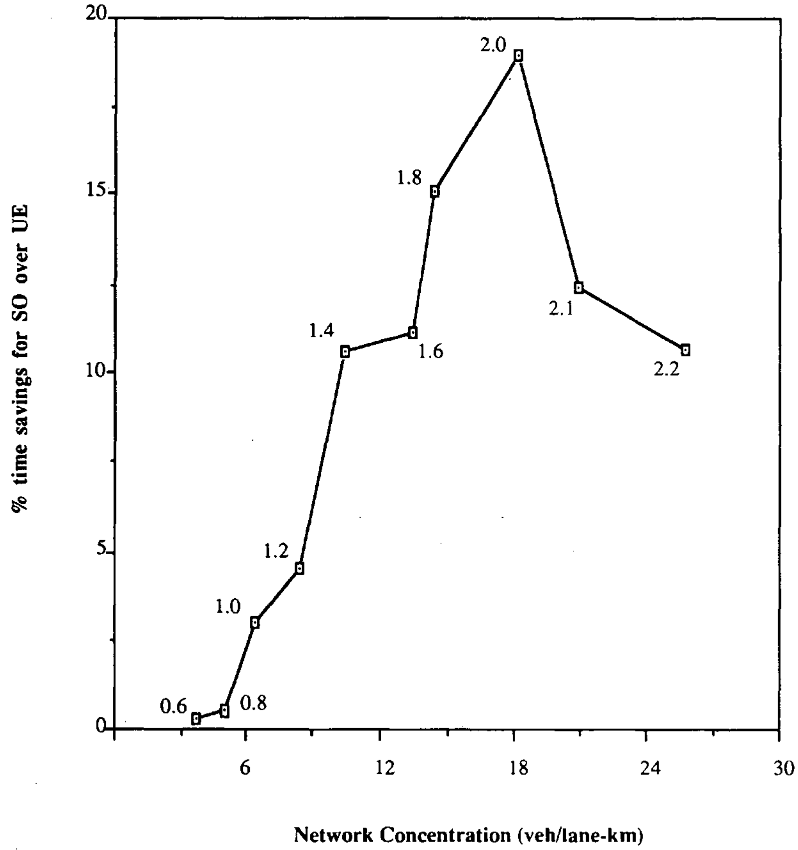


FIGURE 4 Percentage total trip time savings of SO over UE obtained as a fraction of total UE trip time for various loading factors versus average network concentration. The number by each plotted point is the corresponding loading factor.

various levels of network loading. At low loading levels, SO and UE provide essentially identical solutions. For loading factors 0.6 and 0.8, SO shows improvements of 0.3 and 0.5 percent, respectively, over UE. At such low concentration levels, average link speeds remain relatively unchanged because of limited interactions among vehicles, and the marginal travel time on the link is essentially identical to the average travel time, leading to almost identical solutions under the two assignment schemes. When network congestion increases slightly to loading factors of 1.0 and 1.2, the corresponding SO trip time improvements are 3.0 and 4.5 percent, respectively, over the UE solution. As the network becomes moderately congested, system benefits under the SO assignment become more pronounced, with 10.6 and 11.2 percent improvements over UE for loading factors of 1.4 and 1.6, respectively. For heavily loaded networks, very substantial gains are obtained, with 15.1 and 19.0 percent improvements in system travel times using SO for loading factors 1.8 and 2.0, respectively.

As the levels of network loading are increased further, the system reaches very high levels of congestion that near gridlock, and overall network throughput drops, making it increasingly difficult to discharge all vehicles from the system in a reasonable amount of time. Under these conditions, the ability to improve overall conditions by rerouting certain vehicles to paths with lower marginal costs diminishes as all links become highly congested. Thus, the advantage of an SO assignment relative to UE begins decreasing, as reflected by reduced improvements of 12.4 and 10.7 percent for loading factors of 2.1 and 2.2, respectively. The gains begin dropping rapidly beyond this point, with higher loading levels eventually yielding negligible differences in the quality of the solution provided by the two schemes.

Figure 5 depicts the cumulative demand generation as a function of time under the 2.0 loading factor along with the cumulative discharge curves under the SO and UE assignments. The various points on the plot are obtained by accumulating the statistics available for each 5-min interval. The area on the plot between the two discharge curves represents the time savings of SO over UE—in this case about 1,438 hr. The figure illustrates the time-dependent nature of the benefits generated by SO over UE. When the network is in the early stages of loading (for about the first 20 min), it is not sufficiently congested to produce meaningful differences between SO and UE assignments. Most of the savings of SO are accrued between 30 and 70 min into the peak period as the network is close to peak congestion levels. Beyond 70 min, there appear to be virtually no significant gains of SO over UE as the network is again relatively uncongested. Thus the benefits of route guidance based on SO assignment over UE routing are not accumulated uniformly over time—rather they are gained when the network is relatively well congested.

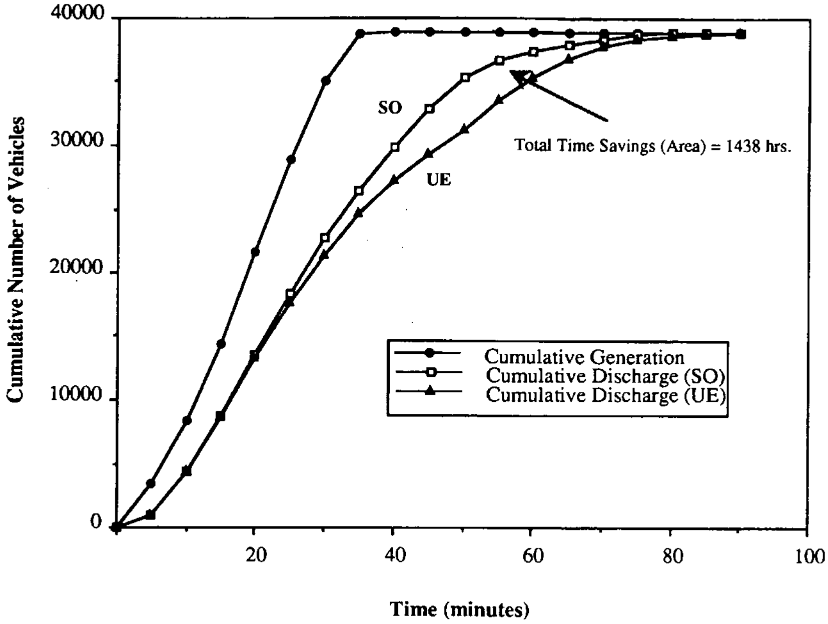


FIGURE 5 Cumulative generation curve and SO and UE cumulative discharge curves for a loading factor of 2.0. The points on the curve represent 5-min updates of the cumulative number of vehicles. The area between the SO and UE discharge curves represents the time savings for SO over UE.

Figure 6 shows the time savings per vehicle for SO over UE as a function of the vehicle's time of departure under various loading factors. To capture the time dependency of the benefits in a systematic manner, travel time savings are accumulated on the basis of the start times of the vehicles. In the figure, 0–5 on the *y*-axis (start time) refers to all vehicles that start between 0 and 5 min. Vehicles that start during the first 5 min do not face congested conditions, and hence SO does not yield savings over UE for these vehicles. Vehicles that start during the intervals 10–15 and 15–20 min accrue time savings at an increasing rate as the loading level increases. Over their trip, these vehicles encounter significant congestion that increases with the loading factor. For vehicles starting between 20 and 35 min, the benefits increase with

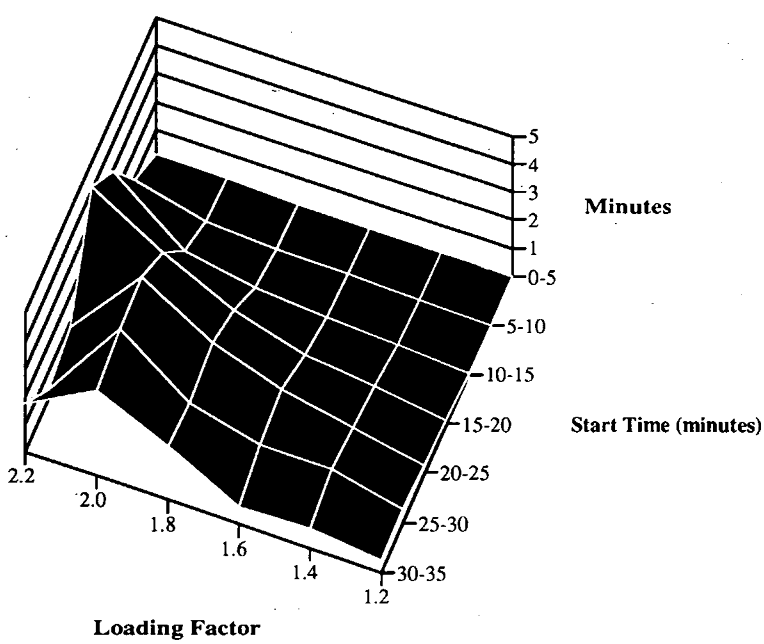


FIGURE 6 Trip time savings (SO relative to UE) per vehicle (in minutes) as a function of loading factor and start times (in minutes) of vehicles.

network loading at an increasing rate until the 2.0 loading factor level and then dip down. This trend illustrates the previously discussed tendency of diminished savings for SO under extremely high congestion conditions.

Network Flow Relations

The second aspect investigated through the experimental results relates to the macroscopic network-level traffic theoretic relationships among network-wide traffic descriptors for dynamic traffic networks under consideration. The pertinent traffic variables and their averages over time and space were defined in the first section of the paper. As noted, although mathematical relationships among traffic flow variables are reasonably well established for arterials and intersections, the intricacies of interactions at the network level preclude analytic derivability of network-wide traffic relationships from the link-level traffic models. However, the simulation results extend the previous findings of Mahmassani et al. (*4,5*) that the basic trends captured by the single roadway relationships seem to also hold at the network level for the dynamic case.

The network level speed–concentration relationship for the SO assignment is shown in Figure 7. Each point on the plot corresponds to a simulation run for the whole assignment period under a particular loading level. The figure clearly illustrates decreasing average network speed with increasing network concentration, paralleling the *K-V* relationship for an individual roadway. Note that the plot has a point of inflection that corresponds approximately to the 1.8 loading factor. This qualitative trend has been observed previously in the simulation experiments of Mahmassani et al. (*4*) on a regular test network using the NETSIM package.

An essential element to be noted in the network-level analysis is the time-dependent nature of the phenomena of interest. Averaging quantities such as network flow and concentration over the duration of the peak period are likely to mask the time dependency of network performance. For example, overall network concentration is obtained by averaging low levels of concentration at both ends of the peak period and high levels in between, as shown in Figure 8, which shows the time-dependent variation of concentration [normalized by dividing by a jam concentration of 96 vehicles per lane-km (160 vehicles per lane-mile)] over the duration of interest. More detailed investigation of the interrelationships among network-level traffic descriptors over time will be reported elsewhere.

CONCLUSIONS

The experiments performed using the simulation-based algorithm to solve both the SO and UE versions of the time-dependent traffic assignment problem have provided insights of critical importance to the design of ATIS information supply strategies and results of fundamental significance in the context of network assignment and network traffic flow theories. Of course, experimental results from a single-test network preclude definitive generalizations; nevertheless, they offer an illustration of the insights that can be obtained on the basic constitution of the problems being addressed while suggesting directions for future research. The first main conclusion is that the results suggest meaningful differences in overall system cost and performance between time-dependent SO and UE assignments. The second main conclusion is that traffic networks under time-dependent traffic assignment patterns continue to operate within the envelope of relatively simple network traffic flow relationships that exhibit strong similarities to the traffic models established for individual road sections.

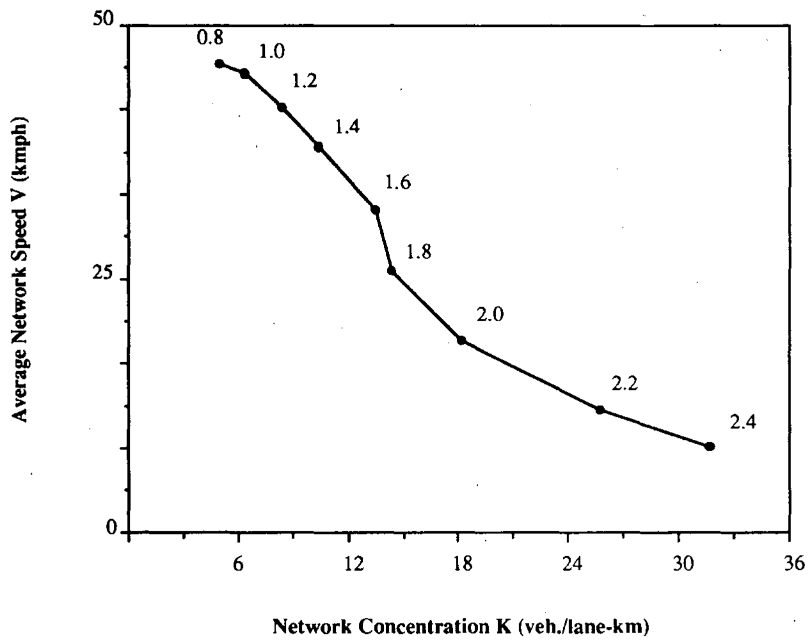


FIGURE 7 Average network speed *V* (kilometers per hour) as a function of average network concentration *K* (vehicles per lane kilometer) for the SO case.

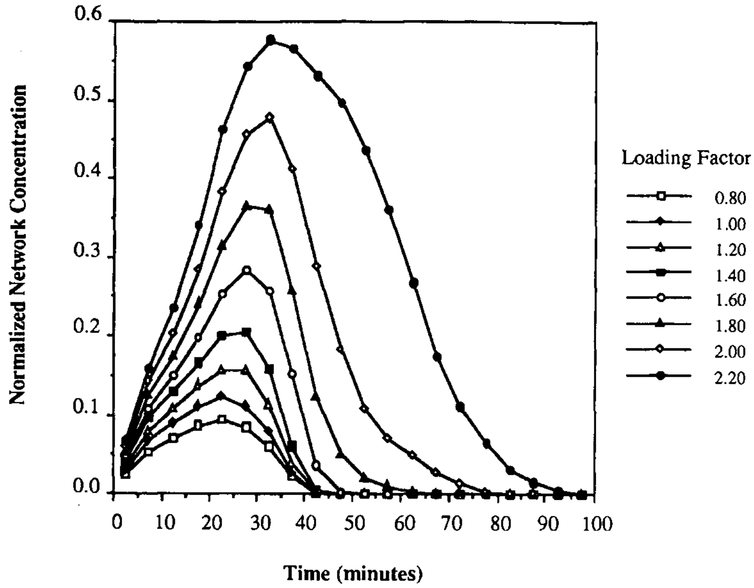


FIGURE 8 Normalized network concentration (network concentration as a fraction of network jam concentration) as a function of time for various loading factors.

If we take the UE assignment results as somehow indicative of the situation that might be attained over time in a system in which drivers have access to real-time on-board descriptive information through ATIS, the results of our experiments suggest that there is considerable potential for SO, coordinated route guidance, especially in heavily congested (although not oversaturated) networks. These results appear to contradict unsupported claims that descriptive information would likely perform as well as normative SO route guidance because UE system costs were claimed to be very close to SO costs. Instead, they strengthen previous recommendations [e.g., by Mahmassani and Jayakrishnan (*9*)] that coordinated information is necessary beyond a certain market penetration level.

The results further highlight the dynamic nature of the benefits accumulated by an SO assignment over UE. They suggest that SO is most effective when the traffic network is moderately to highly congested. In the context of peak period traffic, this implies that most savings through SO assignment would be achieved neither at the beginning nor end of the peak period but in a time range in between. When the network is lightly or very highly congested (oversaturated), an SO assignment does not perform significantly better than UE. For relatively uncongested traffic situations, SO and UE yield almost identical solutions.

Future research on this topic will investigate the system performance under partial user compliance when users are provided with SO paths, thereby introducing an additional element of user behavior. With regard to the traffic network flow theoretic aspects, avenues for future efforts in this area include analyzing dynamic traffic networks from the perspective of the two-fluid theory of town traffic developed by Herman and Prigogine (*10*).

In conclusion, it is possible to characterize traffic flow in urban traffic systems using relatively simple macroscopic relationships that parallel traffic flow relationships at the individual roadway level. Simulation is an abstract representation of real-world traffic; thus the research is mostly exploratory rather than definitive in nature. Results to date strongly suggest that the performance of dynamic traffic networks is critically sensitive to network topology and network loading pattern.

ACKNOWLEDGMENTS

This work is based on funding provided by FHWA under a contract entitled Traffic Modeling To Support Advanced Driver Information Systems. Additional support was provided by the U.S. Department of Transportation through the Southwest Region University Transportation Center.

The authors are grateful to Ta-Yin Hu and Athanasios Ziliaskopoulos for their essential contributions and to Richard Rothery for his beneficial comments and suggestions.

REFERENCES

1. H. S. Mahmassani and S. Peeta. System Optimal Dynamic Assignment for Electronic Route Guidance in a Congested Traffic Network. *Proc., 2nd International Capri Seminar on Urban Traffic Networks*, Capri, Italy, July 1992.
2. H. S. Mahmassani, S. Peeta, G. Chang, and T. Junchaya. *A Review of Dynamic Assignment and Traffic Simulation Models for ATIS/ATMS Applications.* Technical Report DTFH61-90-R-00074-1. Center for Transportation Research. The University of Texas at Austin, April 1992.
3. M. O. Ghali and M. J. Smith. New Dynamic Model to Evaluate the Performance of Urban Traffic Control Systems and Route Guidance Strategies. In *Transportation Research Record 1306*, TRB, National Research Council, Washington, D.C. pp. 33–39, 1991.

4. H. S. Mahmassani, J. C. Williams, and R. Herman. Investigation of Network-Level Traffic Flow Relationships: Some Simulation Results. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., pp. 121–130, 1984.
5. H. S. Mahmassani, J. C. Williams, and R. Herman. Performance of Urban Transportation Networks. In *Transportation and Traffic Theory* (N. H. Gartner and N. H. M. Wilson, eds.), Elsevier Science Publishing Co., Inc., 1987.
6. H. S. Mahmassani, T. Hu, and R. Jayakrishnan. Dynamic Traffic Assignment and Simulation for Advanced Network Informatics (DYNASMART). *Proc., 2nd International Capri Seminar on Urban Traffic Networks*, Capri, Italy, July 1992.
7. A. Ziliaskopoulos and H. S. Mahmassani. Design and Implementation of a Shortest Path Algorithm with Time-Dependent Arc Costs. *Proc., 5th Advanced Technology Conference*, Washington, D.C., Dec. 1992.
8. Y. Sheffi and W. Powell. An Algorithm for the Equilibrium Assignment Problem with Random Link Times. *Networks*, Vol. 12, pp. 191–207, 1982.
9. H. S. Mahmassani and R. Jayakrishnan. System Performance and User Response under Real-Time Information in a Congested Traffic Corridor. *Transportation Research 25A*, pp. 293–308, 1991.
10. R. Herman and I. Prigogine. A Two-Fluid Approach to Town Traffic. *Science*, Vol. 204, pp. 148–151, 1979.

The results and views expressed in this paper are solely the responsibility of the authors and do not necessarily reflect the sponsors' official opinion.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Time-Dependent, Shortest-Path Algorithm for Real-Time Intelligent Vehicle Highway System Applications

ATHANASIOS K. ZILIASKOPOULOS AND HANI S. MAHMASSANI

An algorithm is introduced that calculates the time-dependent shortest paths from all nodes in a network to a given destination node for every time step over a given time horizon in a network with time-dependent arc costs. Unlike other time-dependent algorithms, this approach can handle networks where the travel cost is not necessarily the travel time itself. The algorithm is based on the general Bellman's principle of optimality. It discretizes the horizon of interest into small time intervals. Starting from the destination node, it calculates the paths operating backwards. A proof of the correctness of the proposed algorithm is presented. The algorithm is efficiently implemented and coded on a CRAY Y/MP-8 supercomputer and tested on a large actual street network as well as several random networks. The motivation for this study was the need to compute time-dependent shortest paths in a real-time environment in connection with intelligent vehicle highway systems. The suitability of the proposed algorithm for such applications is demonstrated.

The development of intelligent vehicle highway systems (IVHS) has brought renewed interest in the subject of shortest-path algorithms. For an IVHS control system to respond to rapidly changing traffic conditions in urban street networks, it must be able to calculate optimum routes dynamically. In particular, the calculation of shortest paths in networks with time-dependent arc costs is needed for the dynamic assignment of traffic by a central controller seeking to optimize overall system objectives. This calculation will most likely be repeated a number of times, as part of a real-time decision system, thereby requiring an efficient implementation of a fast dynamic shortest-path algorithm (*1,2*). In addition, the calculation of time-dependent shortest paths will be needed in other parts of the overall system. For example, routines that calculate one-to-all dynamic optimum paths are necessary in simulation models. In addition, on-board computers must be able to calculate one-to-one dynamic shortest paths for the individual needs of the driver.

The above requirements of IVHS control systems provided the primary motivation for this study. To our knowledge, no algorithm in the literature can be implemented in a way that fulfills all of the above requirements. Time-dependent, shortest-path problems are also encountered in a variety of applications in logistics and distribution. This study attempts to fill this gap by presenting and implementing a time-dependent, shortest-path algorithm in a manner that allows it to efficiently calculate paths for large street networks on commercially available computers. The period of interest (e.g., peak period) is discretized into very small intervals. Working in a label-correcting fashion, the algorithm calculates for every interval the time-dependent shortest paths from all the nodes in a network to a given destination node. The algorithm is implemented and coded on a CRAY Y-MP/8 supercomputer and tested on real streets as well as large random networks.

Department of Civil Engineering, The University of Texas at Austin, ECJ Hall 6.204, Austin, Tex. 78712-1076.

LITERATURE REVIEW AND EVALUATION

The first paper dealing with the time-dependent, shortest-path algorithms appears to be by Cooke and Halsey (*3*). These authors developed an iterative function, which is an extension of Bellman's principle of optimality (*4*), that gives the time-dependent shortest paths from every node in the network to one destination node for a set of discrete departure time steps. The travel times on the arcs are defined in multiples of a positive unit of time δ for every time step of the discrete scale $S_M = \{t_0, t_0 + \delta, t_0 + 2\delta, \ldots, t_0 + M\delta\}$. The integer number M is chosen so that the travel times are defined for any $t \in S_M$. The travel times for $t > t_0 + M\delta$ are assumed to be infinite. This assumption eliminates all paths with arrival time to a destination node beyond $t_0 + M\delta$, leading possibly to undetermined paths for some nodes and time steps. This algorithm has theoretical computational complexity $O(V^3M)$, where V is the number of vertexes in the network. However, no implementation scheme for this approach has been reported and hence no computational results are available to determine its actual performance.

Dreyfus (*5*) proposed a label setting approach that generalizes Dijkstra's (*6*) static shortest-path algorithm. This approach calculates the time-dependent shortest path between two nodes for one departure time step with the same computational effort as for the static case [$O(V^2)$]. However, if the paths from all the nodes to a destination node are sought, and for every time step, this approach has the same complexity as Cooke and Halsey's (*3*) algorithm.

An implicit assumption in the Dreyfus approach is that the first-in-first-out (FIFO) property holds on the links of the network. If this assumption does not hold, then the Dreyfus algorithm fails to detect the shortest paths. That was stated in some fashion by several authors, such as Kaufman and Smith (*2*), Halpern and Priess (*7*), Malandraki (*8*), and Orda and Rom (*9*). Orda and Rom recently proposed an approach that is not restricted to FIFO links only. This approach can identify optimum waiting times on the visited nodes when such a waiting is allowed, or the optimum waiting time in the source node if waiting everywhere else is disallowed. However, their approach fails to efficiently find the best path if waiting is not allowed everywhere along the path.

DESCRIPTION OF THE ALGORITHM

Formulation of the Problem

Let $G = (V,E)$ be a V node finite directed graph with E directed edges connecting the nodes. Let $d_{ij}(t)$ be the nonnegative time required to travel from Node i to Node j when departure time from Node i is t; $d_{ij}(t)$ is a real-valued function defined for every $t \in S$, where $S = \{t_0, t_0 + \delta, t_0 + 2\delta, \ldots t_0 + M\delta\}$, t_0 is the earliest possible departure time from any origin node in the network, δ is a small time interval during which some perceptible change in traffic conditions may occur, and M is a large integer number such that the interval from t_0 to $t_0 + M\delta$ is the period of interest (e.g., the traffic peak period).

It is assumed that $d_{ij}(t)$ for $t > t_0 + M\delta$ is constant and equal to $d_{ij}(t_0 + M\delta)$. This is a reasonable assumption for urban transportation networks where, after the peak hour, somewhat stable travel times can be assumed. Nevertheless, it is not a restrictive assumption because M is user defined and can always be increased to include periods with variable travel times on some arcs. It is also assumed that $d_{ij}(\tau) = d_{ij}(t_0 + k\delta)$ for every τ in the interval $t_0 + k\delta < \tau < t_0 + (k + 1)\delta$. This is not a restrictive assumption, considering that by definition δ is very small. Node N denotes the destination node of interest in the network. The algorithm proposed in this paper calculates the time-dependent shortest paths from every node i in the network and at every time step t to the destination node N.

At each step of the computation, denote by $\lambda_i(t)$ the total travel time of the current shortest path from Node i to Node N at time t. Let $\Lambda_i = [\lambda_i(t_0), \lambda_i(t_0 + \delta), \ldots, \lambda_i(t_0 + M\delta)]$ be an M-vector label that contains all the labels $\lambda_i(t)$ for every time step $t \in S$ for Node i. Every finite label $\lambda_i(t)$ from Node i to Node N is identified by the ordered set of nodes $P_i = \{i = n_1, n_2, \ldots, n_m = N\}$.

According to Cooke and Halsey (*3*), $\lambda_i(t)$ is defined by the following functional equation:

$$\lambda_i(t) = \begin{cases} \min_{j \neq i}\{d_{ij}(t) + \lambda_j[t + d_{ij}(t)]\} & \text{for } i = 1,2,\ldots,N-1; t \in S \\ 0 & \text{for } i = N; t \in S \end{cases} \tag{1}$$

A modified version of this equation is the building block of the approach. Instead of scanning all the nodes in every iteration, a list of scan eligible (SE) nodes is maintained, containing the nodes with some potential to improve the labels of at least one other node. The proposed algorithm operates in a label correcting fashion; therefore, the label vectors are just upper bounds to the shortest paths until the algorithm terminates.

The Algorithm

Initially the SE list contains only the destination node N. In the first iteration all the nodes that can directly reach N are updated according to Equation 2 and inserted in the SE list.

$$\lambda_i(t) = d_{iN}(t) + \lambda_N[t + d_{iN}(t)] \qquad i \in \Gamma^{-1}\{N\} \tag{2}$$

where $\Gamma^{-1}\{N\}$ is the set of nodes that can directly reach N. The rest of the labels are set equal to infinity. Next, the first node i of the SE list is scanned according to the following equation:

$$\lambda_j(t) = \min\{\lambda_j(t), d_{ji}(t) + \lambda_i[t + d_{ji}(t)]\} \qquad j \in \Gamma^{-1}\{i\} \tag{3}$$

for every time step $t \in S$. If at least one of the components of Λ_j is modified, Node j is inserted in the SE list. This scheme is repeated until the SE is empty and the algorithm terminates. Relations 2 and 3 are modifications of Equation 1, which in turn is an extension of Bellman's principle of optimality (*4*).

The steps of the algorithm are as follows:

Step 1. Create the SE list and initialize it by inserting into it the destination node N. Initialize the label vectors at the following values: $\Lambda_N = (0, 0, \ldots, 0)$ and $\Lambda_i = (\infty, \infty, \ldots, \infty)$ for $i = 1,2,\ldots,N - 1$.

Step 2. Select the first node i from the SE list, name it "Current Node," and delete it from the list. If the SE list is empty, go to Step 4. Scan the current node i according to Relation 3 by examining each node j, $j \in \Gamma^{-1}\{i\}$. Specifically, for every time Step $t \in S$, check whether $\lambda_j(t)$ is greater than $d_{ji}(t) + \lambda_i[t + d_{ji}(t)]$. If it is, replace $\lambda_j(t)$ in the label vector Λ_j at position i with the new value. If at least one of the M labels of Node j has been improved, insert Node j in the SE list. The details of the structure of the SE list and the associated operations of creation, insertion, and deletion are discussed in the next section.

Step 3. Repeat Step 2.

Step 4. Terminate the algorithm. The M-dimensional vectors Λ_i for every node i in the network contain the travel times of the time-dependent shortest paths from every node i to the destination node N for each time step $t \in S$.

Without proving it, the following theorem is stated: On termination of the algorithm, every element of the vector label is either an infinite number, meaning that no path exists from this node to the destination node at the corresponding time step, or a finite number that represents the shortest path from this node and time step to the destination node.

The proof of this theorem is given by Ziliaskopoulos and Mahmassani (*10*).

IMPLEMENTATION

The implementation of this algorithm is similar to the implementation of a static label correcting algorithm. The three principal implementation issues are the network representation, the data structure of the SE list, and the path storage.

The network representation is more complicated than in the static case because travel times need to be specified for every time step (M steps) for every arc. The most efficient way to store the network is the "backward star" structure because at Step 2 of the algorithm we need all the arcs that end at a specific node. A description of the backward star structure is given by Dial et al. (*11*). To handle the time-dependent trip times, we use the second dimension of the backward star to store pointers to an $E \times M$ matrix, where E is the number of arcs of the network. The required memory to store this structure is the minimum possible, $N + E(M + 2)$ units.

The structure of the SE list for the label correcting algorithms has been studied extensively in the literature (*11–14*). Any SE list structure is appropriate for the proposed algorithm: a simple list with any priority rule, a queue, a double-ended queue, as well as Glover et al.'s partitioning shortest-path scheme with two SE lists (*13*). In this paper, a double-ended queue (deque) structure is implemented. Deque was introduced by D'Esopo and tested by Pallotino (*14*). The deque structure allows the insertion of nodes at both ends of the SE list according to a predetermined strategy and removal from the beginning of the SE list.

The deque is implemented as suggested by Pape (*12*). A one-dimensional array, called deque, holds an integer number and can take the following values:

$$\text{Deque}(i) = \begin{cases} -1 & \text{if Node } i \text{ has been in the SE list at least once but is not there any longer;} \\ 0 & \text{if Node } i \text{ has never been in the SE list;} \\ j & \text{if Node } i \text{ is currently in the SE list and } j \text{ is the node next to it in the list; and} \\ +\infty & \text{if Node } i \text{ is the last node in the SE list.} \end{cases}$$

In addition, two pointers are kept, one pointing to the first (FIRST) and the other to the last (LAST) node in the deque.

We define the following operations associated with this structure:

- Creation: Creation is an initialization step, which is activated just once to set Deque(i) = 0, $i = 1,2, \ldots ,N - 1$ and Deque (N) = ∞. Infinity is defined practically as a very large number—for example, 999,999. This operation also sets the variables FIRST = LAST = N. The whole operation requires $N + 3$ computational time units.
- Insertion: Insertion involves inserting a node at the beginning or the end of the deque. To determine the insertion point, the operation checks the value of Deque(i). If it is 0, indicating that Node i has never been in the deque, the node is inserted at the end of the SE list and the value of the pointer LAST is set equal to i and Deque(i) = ∞. If Deque(i) = -1, Node i is inserted at the end of the deque; Deque(i) is set to FIRST, and the value of FIRST = i. Otherwise, it does nothing because the node is already in the deque. The computational effort required by this step is three time units.
- Deletion: Deletion selects the first element of the deque and assigns it to the variable "CurrentNode." Then, it changes the value of the FIRST to the second element in the deque [which is the Deque(FIRST) node]. It sets the values of Deque(CurrentNode) to -1. The computational effort for this operation is three time units.

The creation operation is called only once from Step 1 of the algorithm and does not contribute significantly to the total computation time of the algorithm. On the other hand, deletion and insertion are called repeatedly from Step 2; as such, they are critical in the determination of the total computational effort of the algorithm.

Finally, the paths are maintained in an $M \times 2$-dimensional array of pointers for each node. These pointers point to the successor node and its label address. This arrangement requires $2VM$ memory locations—the least possible.

In pseudocode form the algorithm is summarized in Figure 1.

The most time-consuming part of the algorithm is Step 2 (see description of the algorithm) in which each element of the M-vector is updated for every node adjacent to the scanned node. This step corresponds to Loop 2 in the pseudocode and requires $4Md$ computational time units, where d is the indegree of the scanned node (number of iterations of Loop 2). Inner Loop 3 can be efficiently vectorized because of the absence of interdependencies, and the number of iterations M is usually greater than 64 (the number of registers in the CRAY's vector processor), which leads to maximum vectorization speed-up (*15*).

The efficiency of the algorithm, however, depends essentially on the total number of scanned nodes before the process terminates. The lower bound on this number is the total number of nodes in the network (V), whereas the upper bound is V^2M. The upper bound is obtained by direct extension of the results for the corresponding static label correcting case. As shown in the next section, this upper bound is not a tight bound in practical applications. The complexity of the algorithm is that of Step 2 (Loop 2) multiplied by the upper bound of the number of repetitions of this step (iteration number of Loop 1) or $O(V^3M^2)$ in the general case that the maximum indegree of a node is $V - 1$.

This implementation of the algorithm was coded in the FORTRAN CFT77 language and run on a CRAY Y-MP/8 supercomputer. The results from the tests are presented in the next section.

COMPUTATIONAL EXPERIENCE

Four different sets of networks are used to test this new algorithm. Set 1 consists of five random networks with a structure similar to that of street networks and with the number of nodes ranging from 100 to 2,500. The number of time steps is held constant at 240. The travel times for each time step are generated in such a way that the FIFO property holds. Specifically, a randomly generated number is accepted as travel time for a given time step only if the absolute value of its difference from the travel time of the previous time step does not exceed the length of the interval between the two steps. Set 1 was designed to test the relation of the performance of the algorithm to the network size.

Set 2 contains five different representations of the same random network consisting of 1,000 nodes, 2,500 arcs, and varying numbers of time steps that range from 120 to 640. In Set 3, the number of arcs ranges from 1,000 to 11,500, whereas the numbers of both nodes and time steps are kept constant at 1,000 and 240, respectively. This set is used to estimate the

```
Call Creation

Call Insertion(N)

Do 1, While (SE list is not Empty)

   Call Deletion(CurrentNode)

   Do 2, For (All nodes J that can directly reach CurrentNode)

      NextNode = J

      InsertInSEList=FALSE

      Do 3, For (t=1,M)

        CurrentTravelTime=TravelTime(NextNode, CurrentNode,t)

        NewLabel=LABEL(CurrentNode,t+CurrentTravelTime)+CurrentTravelTime

        If (LABEL(NextNode,t) ≤NewLabel) Then

           LABEL(NextNode,t)=NewLabel

           InsertInSEList=TRUE

           PathPointer(NextNode,t,1)=NodeCurrent

           PathPointer(NextNode,t,2)=t+CurrentTravelTime

        End If

3         Continue

        If (InsertInSEList) Call Insertion(NextNode)

2       Continue
```

FIGURE 1 Algorithm for program time-dependent shortest path.
(*continued on next page*)

relationship between the execution time and the average degree of a node in a network.

Finally, Set 4 consists of one real street network—that of the core area of Austin, Texas—consisting of 625 nodes and 1,724 arcs. Time-dependent travel times for this network were produced from a simulation model called DYNASMART (Dynamic Network Assignment Simulation Model for Advance Road Telematics), for a simulated peak period of 50.3 min. This peak period is discretized into 503 intervals of 0.1 min each, and the travel time for each interval is generated.

Tables 1 through 4 present the computation times in CPU milliseconds for each set. All the runs were performed on a CRAY Y-MP/8 supercomputer, using the CFT77 FORTRAN compiler. This computer has eight CPUs with vector pipeline architecture. The algorithm is coded to allow vectorization when applicable. Vectorization is especially well suited for Step 2 of the algorithm, in which M iterations are performed because no dependency exists between any two iterations, and the number M is usually larger than the number that CRAY considers the minimum number of iterations for maximum speed-up. However, no attempt was made to exploit other hardware characteristics of the CRAY beyond vectorization. In addition, to smooth out the effect of the destination node choice on the execution time, 30 runs were performed for 30 different destinations for every network, and the average computation time is reported.

Tables 1 and 2 contain the results for Network Sets 1 and 2, which indicate that the computation time increases almost linearly with the number of nodes and the number of time steps in the network.

Table 3, on the other hand, suggests that a nonlinear relationship exists between the execution time and the average degree of a node in the network. An exponential model was calibrated from these data using regression, yielding the following relationship:

$$\text{Computation time} = 22.13\ d^{1.4}$$

where d is the average indegree of a node in the network.

Table 4 contains the averages and standard deviations of the computation time and the total number of scanned nodes for the real street network of the Austin, Texas, core area. The total number of scanned nodes is the main factor that affects the performance of the algorithm. The lower bound for this number is the number of nodes in the network (V), whereas an upper bound was found to be V^2M in the previous section. Table 4 shows that for the tested network of 625 nodes, the total number of scannings was 736, or $1.18V$, which is considerably less than the theoretical upper bound. Moreover, from the low values of the standard deviations, it can be inferred that the algorithm is reasonably stable.

Combining all the above results, we can conclude that as is common with shortest-path problems,. the actual computational performance for the networks considered here is on the order of $VMd^{1.4}$, which is far from the worst-case theoretical complexity $O(V^3M^2)$. In addition, as mentioned ear-

```
1       Continue

        Procedure Creation
          Do, For (Node=1, V-1) Deque(Node)=0
          Deque(N)=999999
          FIRST=N
          LAST=N

        Procedure Deletion(CurrentNode)
          CurrentNode=FIRST
          FIRST=Deque(CurrentNode)
          Deque(CurrentNode)=-1

        Procedure Insertion(Node)
          If (Deque(Node)=0) Then
            Deque(LAST)=Node
            LAST=Node
            Deque(Node)=999999
          Else
            If (Deque(Node)=-1) Then
              Deque(Node)=FIRST
              FIRST=Node
            End If
          EndIf
```

Where:

LABEL(Node,t) is a variable that holds the M-vector labels for every node.

PathPointer(Node,t,1) is a pointer that points to previous node while PathPointer(Node,t,2) points to the corresponding time of arrival at the previous node of the shortest path from this node to the destination node N.

InsertInSEList is a logical variable which is used to determine if a label of a node was changed.

NewLabel is an auxiliary variable that temporarily holds the new label of the next node.

FIGURE 1 ***(continued)***

lier, the algorithm was vectorized efficiently. The algorithm was tested on the Austin core street network with the vectorization feature disabled for the same destination nodes as above, and the average execution time was found to be 728.02 msec. This means that the vectorization in this case yielded a speed-up of 6.74 (speed-up is defined as the ratio of the total computation time of the algorithm without vectorization to the corresponding computation time with vectorization).

Next, we compare our approach with the expanded static case proposed by Dreyfus (*5*). We implemented that scheme as efficiently as we could and achieved an execution time of 2.2 msec to find the time-dependent shortest path between

TABLE 1 Computation Times (msec) for Various Network Sizes

Network	Nodes	Arcs	Time intervals	Comp. time
1	100	250	240	5.97
2	500	1250	240	35.73
3	1000	2500	240	73.28
4	1500	5000	240	141.42
5	2500	8000	240	235.04

TABLE 2 Computation Times (msec) for Various Numbers of Intervals

Network	Nodes	Arcs	Time intervals	Comp. time
1	1000	2500	120	37.95
2	1000	2500	240	73.28
3	1000	2500	360	102.46
4	1000	2500	480	131.56
5	1000	2500	640	158.54

TABLE 3 Computation Times (msec) for Various Numbers of Arcs

Network	Nodes	Arcs	Time intervals	Comp. time
1	1000	2000	240	55.89
2	1000	3000	240	91.88
3	1000	6000	240	253.95
4	1000	9000	240	448.19
5	1000	11500	240	624.86

TABLE 4 Performance of the Algorithm on Part of Austin's Core Street Network Consisting of 625 Nodes and 1,724 Arcs for 503 Intervals

	Computation time in milliseconds	Total number of scanned nodes
Mean	107.41	736
Standard Deviation	11.82	81

one origin and one destination for one time step on the Austin core network. To compare it with our proposed algorithm, the time-dependent shortest paths must be calculated from all 625 nodes of the network to one destination and for 503 time steps for each node. This calculation would require a total time of 0.0022625503 = 691.25 sec. However, the calculation of one path for one node and one time step produces at the same time the paths to the destination from every node along the path for one time step. The maximum number of nodes in a path for the real street network was 72. Therefore, we can estimate a lower bound on the total execution time by assuming that every time one path is computed, 72 other not previously calculated paths are obtained at the same time. This lower bound is 9.6 sec, which greatly exceeds the 0.107 sec achieved with the proposed algorithm.

The proposed algorithm takes advantage of two main characteristics of networks with time-dependent arcs. One is that only a few paths between a given OD pair become best paths at any point in time. Usually, three or four paths are interchanged as best paths at different time steps, with one path often maintaining its best path status for most of the time. For example, the maximum number of paths observed during the testing of the real street network was 17 (out of a possible 503).

The second characteristic of dynamic networks is that even if various paths were best at different times between a given OD, these paths would be likely to share the same next-to-the-origin node (i.e., second node in the path). This means that most of the best paths from a given node result from the scanning of just one of the neighboring nodes. The effectiveness of our algorithm is attributed to these two reasons. Specifically, the fewer the paths that are best at different times, the closer is the behavior of the algorithm to that of static label correcting algorithms. In the extreme case that the same path remained best between a given OD pair for all the time steps, the corresponding origin node would contribute to the total computation time of the algorithm as if the network were static. Even if more than one path were best for a given OD pair, these paths could be calculated in just one scanning of a neighbor node. From Table 4, we can see that for the real street network of Austin, Texas, only 111 (736 − 625) nodes were scanned for a second time, although in general different paths were best at different times for a given origin node.

Finally, the proposed algorithm does not require the FIFO that the Dreyfus approach requires because it operates in a label correcting fashion. That makes the algorithm applicable to networks with time-dependent arc costs that arise in a variety of areas, not just transportation networks. Examples include equipment replacement policy, vehicle routing and scheduling, capacity planning, and communication networks.

SUMMARY

In this paper, a new time-dependent shortest-path algorithm was introduced. It calculates simultaneously all the shortest paths from all nodes to a given destination node and for every discrete time step in a network with time-dependent arc costs. The algorithm is based on Bellman's general principle of optimality and can be applied to any network with time-dependent arc costs. The correctness of the algorithm was analytically established.

An implementation scheme for the algorithm was proposed, coded, and run on a CRAY Y-MP/8 supercomputer. The coded scheme was tested on a set of random networks and one real street network. The computational results demonstrated the efficient performance of the proposed algorithm for a broad range of network structures. This efficiency is attributed to the fact that the algorithm is not just an extension of a static algorithm but is designed to take advantage of the specific characteristics of networks with time-dependent arc costs.

ACKNOWLEDGMENTS

This paper was based on work funded by FHWA, U.S. Department of Transportation, through a contract entitled Traffic Modeling to Support Advanced Driver Information Systems. The computing time and the technical support provided from the Center for High Performance Computing at the University of Texas at Austin are also gratefully acknowledged.

The authors appreciate the discussions with the other project researchers, especially Ta-Yin Hu, Srinivas Peeta, and Richard Rothery.

REFERENCES

1. H. S. Mahmassani, G. L. Chang, S. Peeta, and T. Junchaya. *A Review of Dynamic Assignment and Traffic Simulation Models for ATIS/ATMS Applications*. Technical Report DTFH61-90-R-00074-1. CTR, The University of Texas at Austin, 1992.
2. D. E. Kaufman and R. L. Smith. Minimum Travel Time Paths in Dynamic Networks with Application to Intelligent Vehicle/Highway Systems. *IVHS Journal*, in press.
3. K. L. Cooke and E. Halsey. The Shortest Route Through a Network with Time-Dependent Internodal Transit Times. *Journal of Math. Anal. Appl.*, Vol. 14, 1966. pp. 492–498.
4. R. Bellman. On a Routing Problem. *Quart. Appl. Mathematics*, Vol. 16, 1958, pp. 87–90.
5. S. E. Dreyfus. An Appraisal of Some Shortest-Path Algorithms. *Operations Research*, Vol. 17, 1969, pp. 395–412.
6. E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numer. Mathematics*, Vol. 1, 1959, pp. 269–271.
7. J. Halpern and I. Priess. Shortest Path with Time Constraints on Movement and Parking. *Networks*, Vol.4, 1974, pp. 241–253.
8. C. Malandraki. *Time Dependent Vehicle Routing Problems: Formulations, Solution Algorithms and Computational Experiments*. Ph.D. dissertation, Northwestern University, Evanston, Ill., 1989.
9. A. Orda and R. Rom. Shortest-Path and Minimum-Delay Algorithms in Networks with Time-Dependent Edge-Length. *Journal of the ACM*, Vol. 37, 1990, pp. 607–625.
10. A. K. Ziliaskopoulos and H. S. Mahmassani. Design and Implementation of a Shortest Path Algorithm with Time-Dependent Arc Costs. *Proc., of 5th Advanced Technology Conference*, Washington, D.C., 1992, pp. 1072–1093.
11. R. B. Dial, F. Glover, D. Karney, and D. Klingman. A Computational Analysis of Alternative Algorithms and Labeling Techniques for Finding Shortest Path Trees. *Networks*, Vol. 9, 1979, pp. 215–248.
12. U. Pape. Implementation and Efficiency of Moore-Algorithms for the Shortest Route Problem. *Mathematical Programming*, Vol. 7, 1974, pp. 212–222.
13. F. Glover, R. Glover, and D. Klingman. The Threshold Shortest Path Algorithm. *Networks*, Vol. 14, 1986, pp. 132–145.
14. S. Pallotino. Shortest-Path Methods: Complexity, Interrelations, and New Propositions. *Networks*, Vol. 14, 1984, pp. 257–267.
15. *Parallel Processing Guide. CF77*, Vol. 4. Cray Research Inc., Mendota Heights, Minn., 1991.

The contents of the paper are the sole responsibility of the authors.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Communications Architecture for Early Implementation of Intelligent Vehicle Highway Systems

D. J. CHADWICK, V. M. PATEL, AND L. G. SAXTON

Communications—wireless communications in particular—is a critical component of intelligent vehicle highway systems (IVHS). It is costly when viewed from two different angles: first, its dependence on using the scarce natural resources called radio frequency (RF) spectrum, and second, the actual cost of implementing the necessary infrastructure and the required (in-vehicle) user equipment. On the other hand, it is imperative, for the success of the national IVHS program, to plan a near-term IVHS architectural implementation to show positive first-user benefits. Bearing in mind the above constraints, it is natural to think of designing an IVHS communications architecture that makes use of existing infrastructures for other communications services. This strategy enables more efficient use of the RF spectrum while it reduces the total cost of services by sharing the communications infrastructure and end-user equipment. A communications architecture is proposed for IVHS called the Subsidiary Communications Authority Traffic Information Channel (STIC), based on the widely available FM radio broadcast services' infrastructures by making use of FM subcarrier technology. This preliminary design also shows that STIC has a higher data transmission capacity than any other existing FM subcarrier broadcast system and that it has the potential to meet the one-way outbound (broadcast) data transmission capacity needs of IVHS for the next few years. In addition, STIC architecture is capable of being scaled up in the future.

The intelligent vehicle highway system (IVHS) program is broadly described as applying modern communication and control technology to the needs of highway transportation. In this regard, IVHS is yet another part of society that is increasingly interwoven with and dependent on modern communications technology. At home, at the business place, or in transit, currently available communications provides an additional dimension of information exchange to enhance one's business and personal life.

The success and viability of IVHS is dependent on several factors; however, perhaps none will be more central and fundamental than efficient, reliable, and affordable communications. Although some IVHS functions will be serviced by "hard-wire"-type communications, the more advanced system concepts generally involve communication with a moving vehicle and thus require some form of wireless approach. At issue here is the substantial growing demand for RF spectrum to support various perceived and developing markets such as high-definition television and advanced personal communications. Given these spectrum constraints, there is a real incentive to overlay IVHS communication functions on existing services where possible. Two examples that have been discussed are digital cellular and FM subcarrier.

IVHS is not a single system but rather a broad set of applications with functions and technology that intersect in various situations. IVHS also has a range of technology and system availability spanning products that are available today to concepts not expected to be marketed until after the turn of the century. This range of both product types and market timing poses a dilemma to IVHS. For the most efficient system and effective use of spectrum there is the need to establish an overall efficient architecture for IVHS. Efforts to make this analysis and to postulate the most preferred architecture are now being initiated but will require at least 2 to 3 years. However, there is an immediate need to bring to market and thus realize the benefits of many IVHS applications that are near-term implementation candidates. A primary example is the desire to market motorist information and route guidance-type systems. Thus, this early market need results in the necessity to establish communication resources and standards now for these current applications without benefit of knowing the form of the final IVHS architecture. The practical compromise is to attempt to establish a communication design to service these near-term needs that is easily available, minimizes any Federal Communication Commission (FCC) rule-making needs, is economical, and is adaptable to future architecture designs (*1*).

Given these objectives and constraints, FHWA has been considering various communications approaches to service motorist information and route guidance-type applications. These systems do not necessarily have a single or homogeneous set of communication needs as there is a strong need to provide both low-cost, basic performance systems and higher-cost, high-performance systems. Low cost can be characterized as radio data system (RDS)-traffic management system (TMS) type systems, whereas the higher-performance systems are of the TravTek type.

The use of the subcarrier part of each FM station's frequency assignment has some particular advantages—especially for the low-end systems (*2*). However, there also appears to be the opportunity to reconfigure the subcarrier in a manner in which the FM subcarrier could also service high-end systems. This potential has been the focus of recent efforts, and this paper is part of an effort to share this promising communications approach with the IVHS community. If further analysis and preliminary field testing support the value

D. J. Chadwick and V. M. Patel, The MITRE Corporation, 600 Maryland Ave., S.W., Washington, D.C. 20024. L. G. Saxton, FHWA, Turner Fairbank Highway Research Center, McLean, Va. 22101.

of this approach the objective would be to work it into early operational tests.

COMMUNICATIONS OUTLETS FOR EARLY IMPLEMENTATION

Public policy analysts generally agree that new systems such as IVHS must provide their first increment of user services or public benefits within about 5 years of the inception of the project. Further, the initial offerings to the public have to be "winners"; that is, they have to be perceived as providing the benefits that they promised at their inception, at a cost commensurate with those benefits. The Interstate highway system is a good example of this incremental implementation strategy. In this case, the new highways were opened to traffic after the completion of segments only a few miles long, thus delivering their promised benefits (safer, higher-capacity, higher-speed roadways) quickly and in ever-increasing measure.

For IVHS to operate in the same way, it is essential that initial implementations not depend on the design, development, and installation of large infrastructures. This is especially true in the case of communications systems because they have to be in place before any other subsystems can even be tested effectively. This drives the conclusion that early IVHS projects will have to make use of existing communication outlets to the maximum possible extent.

Several available communications technologies are available for early IVHS projects. Each will be discussed in turn, elaborating on issues such as capacity, coverage, and cost.

RDS

RDS is widely used in Europe for a variety of services including traffic information (*3*). Broadcast stations in the United States plan to install the system as well, but its applications may well degenerate to the single function of program type identification (classical, rock, sports, news, etc.) because the system's capacity and hence level of detail for route guidance data transmission are very low. RDS uses an AM subcarrier on the FM broadcast station that is located at 57 kHz on the FM baseband. The total baseband bandwidth that RDS occupies is about 7 kHz or from just above the stereo L-R spectrum limit at 53 kHz to 60 kHz. The gross data rate for the channel is set at 1,187.5 bps, almost half of which is used for error correction. After extracting the capacity used by fixed RDS functions, the data rate remaining for "additional" services is under 300 bps. That rate will not begin to support IVHS data requirements for traffic link time updates, for example. RDS is, however, a robust system. Its cost is also relatively low. FM broadcast stations in the United States plan to adopt the system, if only to support a program-type function.

Highway Advisory Radio and Advanced Highway Advisory Radio

Highway advisory radio (HAR) and advanced highway advisory radio (AHAR) systems are widely used in the United States for broadcasting information to travelers in a limited area. Services include traffic information, parking availability at airports and national parks, scenic view alerts, and so forth. The systems use AM broadcast-band equipment, operating at low power levels. Until recently, the transmitters were operated on fixed frequencies (530 and 1610 kHz), within the frequency range available to standard car radios. In the future, HAR and AHAR stations will be assigned anywhere in the AM broadcast band that is available, considering the assignments to local commercial stations. Because they operate in a band in which the maximum modulating frequency is 5 kHz, the capacity of the systems to handle data is limited and is meant to be a voice announcement medium. Because of their limited range, HAR stations would have to be installed to support IVHS services in nearly all cases, which goes against the strategy of limiting infrastructure installation initially.

FM/Subsidiary Communications Authorization

As part of its license, every FM broadcast station is granted the authority to broadcast other program material on subcarriers that cannot be detected by standard receivers (see Figure 1). These subsidiary communications authorization (SCA) subcarriers are already used by some stations as sources of extra revenue and carry program material that ranges from sports programs and paging data to background music and stock quotes. Many stations do not currently use their SCA subcarriers. FM broadcast stations provide a high grade of service to well over 90 percent of the area of the United States and to 100 percent of the area where early IVHS experiments are likely to be conducted. The equipment needed to enable the SCA signal is inexpensive, and no modification to the station is required. Likewise, SCA receivers are available at low cost. The capacity of the SCA channel, given suitable

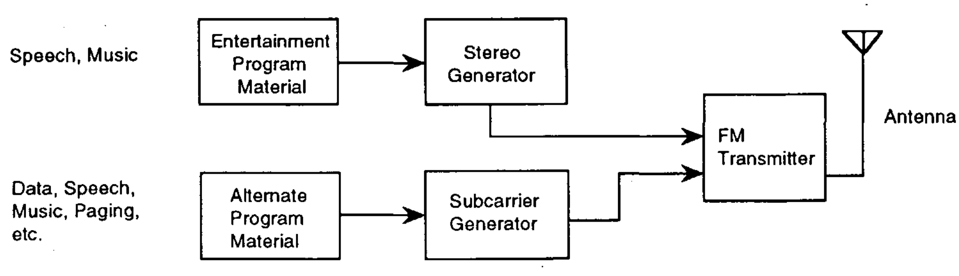


FIGURE 1 FM broadcasts with SCA.

modulation and coding approaches, can be more than 16 kbps, as is discussed later in this paper. Thus, an installed infrastructure of high-capacity, wide-area communications exists to support IVHS needs for communications from the highway to vehicles.

Television Data Channels: Vertical Blanking Interval and Aural Subcarriers

Broadcast TV channels offer two different outlets that will support data communications for IVHS. The first system, called vertical blanking interval (VBI), modulates the first 20 picture lines of each frame, which cannot be "seen" on the screen because they occur during the interval (blanking) when the picture is "off." VBI already is used extensively to provide closed captioning for the aurally handicapped. This service is supported on a single VBI line (Line 4). The technology exists to chain from 2 to all 20 lines into a single virtual data channel, with a resultant capacity of about 56 kbps, although it is somewhat complex. Receivers to use the VBI channel are basically full TV visual implementations because the signals have to be recovered from the composite video baseband. At present the only VBI receivers that are in production are limited to the closed caption service and would not support general data communications.

A TV station is really two "associated" broadcast stations with separate transmitters for the visual and aural signals (Figure 2). The TV aural transmitter is actually just an FM station with somewhat different technical characteristics. As such, subcarriers can be supported as with conventional FM. One of the standard TV aural subcarriers is called the secondary audio program (SAP). It is infrequently used for some of the same services as FM SCA and occasionally to broadcast a running commentary of the main TV program for the visually handicapped. The SAP channel has a much wider bandwidth than the SCA subcarriers, although the basic technology is identical. SAP receivers are available only as a separate device (i.e., not built into a TV set) from one manufacturer at present. Since the bandwidth available on a TV aural signal is large, more than one wideband subcarrier can be accommodated. Even in those cases in which a TV station is using SAP, another subcarrier can be added for traffic data.

Because there are fewer TV stations than FM stations in most areas, and because their coverage area is typically smaller, the TV data channels may be less desirable in some situations. However, in major metropolitan areas where needs warrant, these outlets could provide an extra measure of capacity.

Land Mobile Frequencies in the Range of 220 to 222 MHz

Although it is certain that all IVHS communications requirements cannot be met by any single system, there are a few basic critical functions for which the desirability of continent-wide common frequencies is obvious. To accommodate fundamental services in the safety and warning area, which should be available to all "classes" of users, these messages should be carried on frequencies that are the same everywhere so that the simplest possible radio design can be supported. The National Telecommunications and Information Administration and FCC recently opened a new land mobile band between 220 and 222 MHz. The band plan allots frequencies to several kinds of users over a set of 200 channel pairs, each 5 kHz wide. Besides being divided between government and nongovernment users, the frequencies are allotted to national and regional applications. Within the nongovernment set, users in both commercial and noncommercial classes are accommodated. The FCC has received applications from about 80,000 individuals for nongovernment licenses in the band, but the principal interest by these people is in the commercial frequency pairs.

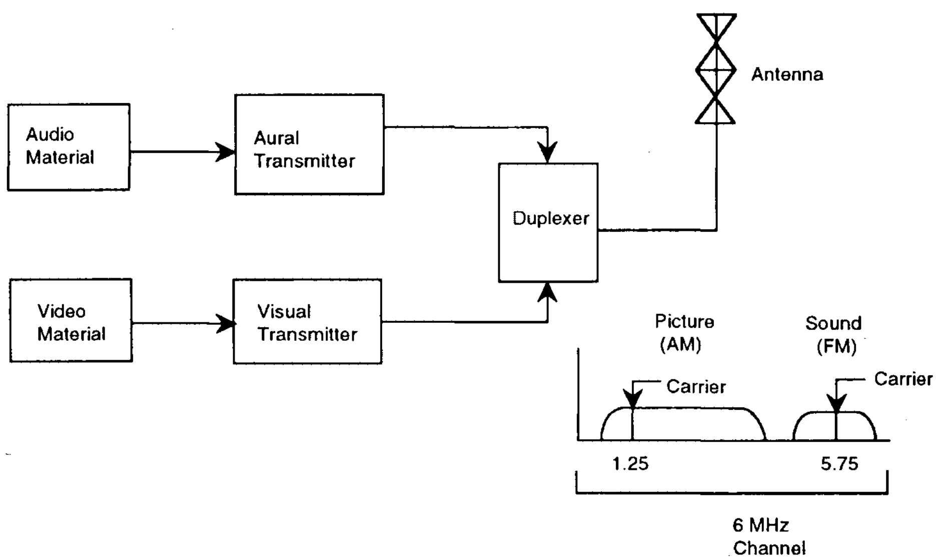


FIGURE 2 Television broadcasting.

Besides the safety and warning services mentioned above, the use of continent-wide frequencies in this band would enable a "hailing" function, wherein the other frequencies in use for other services in a given area can be broadcast and used to automatically tune the in-vehicle receivers. This approach unencumbers the design and frequency acquisition of systems from the absolute need to be exactly the same across the continent.

DESIGN FEATURES OF AN SCA-BASED TRAFFIC INFORMATION CHANNEL

Within the FCC regulations for FM radio broadcast services, the SCA subcarrier can be placed anywhere in the range from 60 to 99 kHz on the baseband (*4*). Figure 3 shows the baseband spectrum of a commercial FM broadcast system with such an arrangement. The subcarrier at 76 kHz can be modulated by the digital traffic data stream and then used to frequency modulate the main carrier to a level corresponding to a modulation index of 10 to 20 percent. Although the FCC limits the injection level of the subcarrier modulation on the main FM carrier to 10 percent when the subcarrier is placed at 76 kHz, the FCC should not object to using an injection level of 10 to 20 percent for STIC. Placing the subcarrier at 76 kHz improves reception and simplifies design because 76 is a fourth harmonic of the pilot tone at 19 kHz. Eventually, this feature reflects as a cost-reducing factor for the mobile receiver.

To reduce the synchronization and timing requirements on the receiver circuitry that demodulates and extracts the SCA traffic information, the digital traffic data rate of the basic modulating signal of the SCA channel is derived via a "binary division" of the subcarrier frequency (76 kHz) itself. A division by eight of the subcarrier frequency would give a base SCA channel baud rate of 9.5 kbps. Using a quadrature phase shift keying (QPSK) modulation scheme for the subcarrier modulation will give a data rate of about 19 kbps (twice the baud rate).

As shown in Figure 3 for the FM/STIC baseband spectrum as well as in Figure 4, which shows the channel structure for

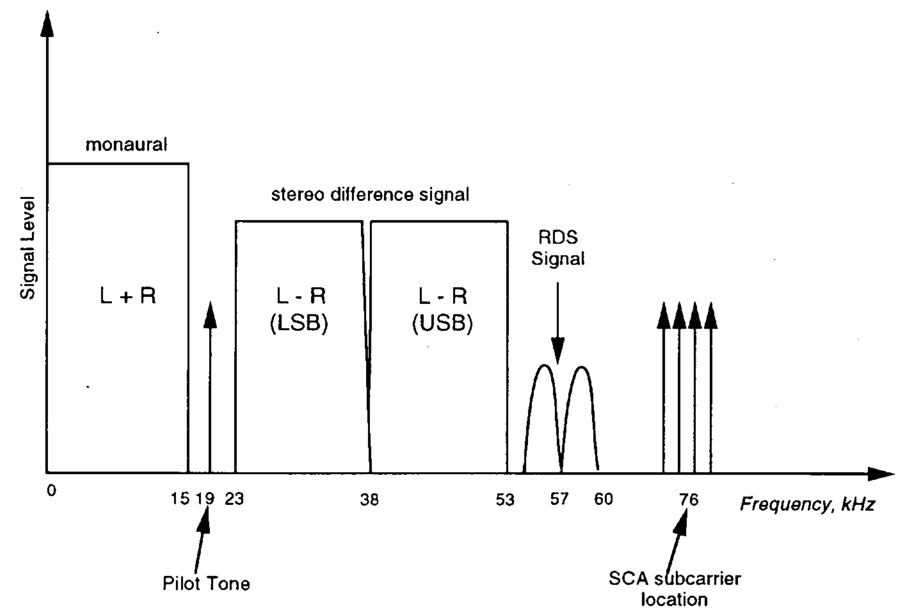


FIGURE 3 FM baseband format under the STIC system.

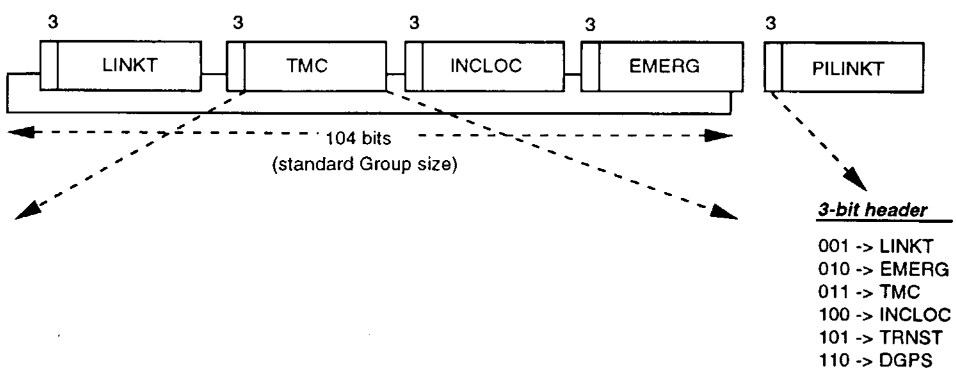


FIGURE 4 SCA STIC structure.

STIC system, the RDS data structure was subsumed as far as possible. The size of the data blocks ("groups" in RDS terminology) is maintained at 104 bits long, as in the European RDS. In addition, a three-bit header is added in front of each data block to signify what type of data block follows. Presently, the following types of data blocks are defined:

1. LINKT data groups will provide the updates for the link times to the vehicles as necessary. They also will be used for periodic refreshing of the basic link-time data base resident in the vehicles (e.g., with a periodicity of once every 5 min). The group structure is shown in Figure 5. This type of message is essentially for use with the high-end family of IVHS services and will be ignored on reception by the in-vehicle STIC system equipment for the low-end services users.
2. EMERG group will broadcast emergency-related data when the situation demands. A severe accident involving toxic chemical spill is a good illustration. Although this data group may be broadcast infrequently, it will be assigned the highest priority and may preempt transmission of any other data group. This group is applicable to both the high- and low-end services users.
3. TMC group is the European RDS-TMC data group (*5*). It is planned to adopt the ALERT "C" protocol, with modifications to suit the U.S. requirements, containing most features as proposed in ALERT "C." This data group again is for both high- and low-end services users.
4. INCLOC: Whenever a TMC group transmits incident-related information that has associated location information to be broadcast, it is sent out via one or more INCLOC blocks. Figure 6 shows the details of the structure of this group. This message group is applicable to both the high- and low-end services users.

The TRNST block conveys desirable schedules of, and connections to, the public transit system (both surface and subsurface systems). The DGPS block may be used to broadcast the correctional (time and location) parameters computed by the differential GPS station for the area.

STIC CAPACITY AND IVHS COMMUNICATIONS LOAD

It is estimated that even with full advanced traveler information system (ATIS)-advanced traffic management system (ATMS) implementation, a downlink gross data rate of about 19 kbps would be sufficient to handle required services, in-

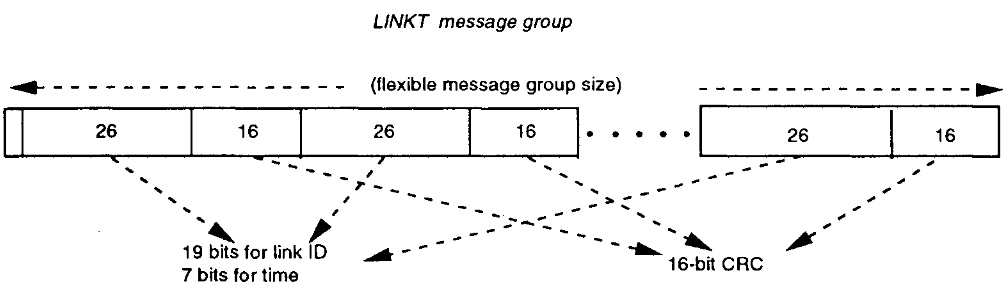


FIGURE 5 STIC group structures (LINKT).

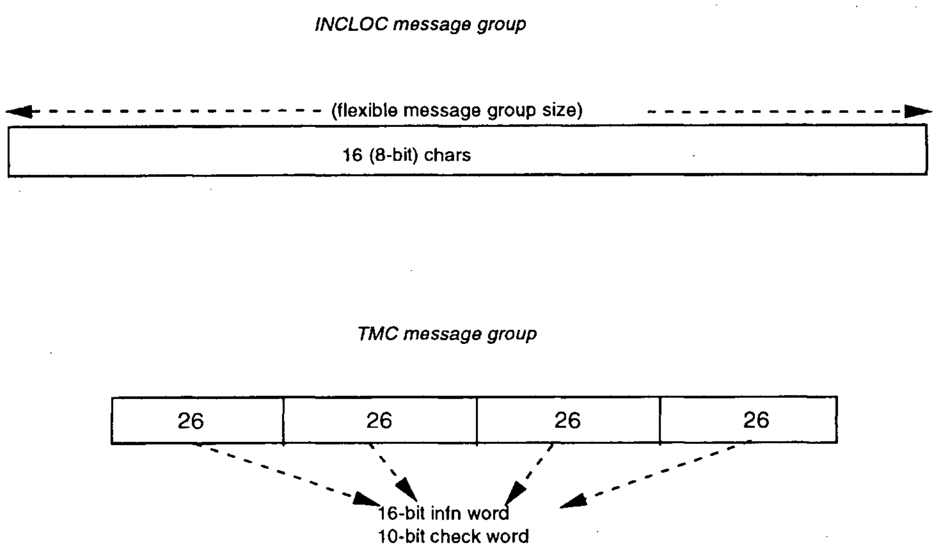


FIGURE 6 STIC group structures (INCLOC and TMC).

cluding updating the in-vehicle link-time data base with acceptable periodicity. Given below are the salient features of the communication downlink load for IVHS along with an example computation.

- INCLOC capacity sample calculation: assuming eight alpha characters for a street and two streets to define an intersection: 8 × 2 = 16 characters for every intersection (abbreviated location information in text form), and no error detection or correction. Using the total STIC capacity of 19 kbps for INCLOC message blocks only, and with 16-character (or 131-bit) blocks, 19,000/131, or about 145 intersection-level location information sets, could be passed on to the motorists each second.
- STIC capacity sample calculation:
  —Assuming the following set of traffic information load: one INCLOC group is necessary to be transmitted for every TMC group, and TMC groups consume 6 percent (1.2 kbps) of the total channel capacity (19 kbps). This means 1.2 kbps are consumed by TMC messages and 1.47 kbps by INCLOC messages. Thus, the STIC capacity available for LINKT groups alone is 19 − 2.67 = 16.33 kbps.
  —Optimally, having a flexible LINKT message group (say, 213 bits long), five link times could be broadcast via one group. Utilizing the total remaining capacity every second: (16,330/213) × 5, or about 383 link times, could be broadcast along with about 11 TMC groups and 11 INCLOC groups. In other words, with about 16 kbps channel capacity allotment to LINKT blocks, more than 22,000 link times may be transmitted each minute—which is perhaps adequate, at least for the initial few years of IVHS deployment.
- QPSK subcarrier modulation will be used to provide double the data rate (19 kbps) from a baud rate of 9.5 kbps. A modulation efficiency of 1.5 bits/Hz is not very sophisticated by today's state-of-the-art modulation standards. Only a modest amount more in system cost could push the practical channel speed further upwards as requirements expand.
- To keep the end equipment relatively inexpensive, data compression is currently not incorporated. However, it may be a viable approach to increase the data rate in the future and should be evaluated.
- Service range: Although the modulation level of the main carrier allowed by the SCA signal is about half that of the main program signal, the SCA modulating signal would be digital in nature and, consequently, can tolerate a lower ratio of signal to noise (SNR) at the receiving end. Computations show that the geographical range of coverage for the STIC information signal would be almost the same as that for the standard FM entertainment signal.
- STIC is scalable to more sophisticated ATIS/ATMS deployment.

ESTIMATION OF SERVICE RANGE FOR STIC

An FM detector, called a discriminator, produces an output voltage that is proportional to the frequency deviation of the input signal. Thus for a given SNR, the range reduction for the SCA channel, with a modulation index of 0.2 (as compared with the main channel whose index is typically 0.35 if SCA is present) can be estimated from the relationship:

$$20\log_{10}(0.2/0.35) = 20\log_{10}(0.57) = 20(-0.24) = -4.9\text{dB} \quad (1)$$

This is the SCA channel output power as compared with that of the main entertainment channel(s).

From the free-space loss equation,

$$Lp = 20\log_{10}(d) + 20\log_{10}(f) + K \quad (2)$$

a 4.9-dB reduction in signal power corresponds to a range reduction of

$$4.9\text{ dB} = 20\log_{10}(d) \quad (3)$$

That is, $\log_{10}(d) = 0.245$, or $d = 1.76$. Hence, the service range of the SCA channel is 1/1.76 or 57 percent that of the main program for the same SNR.

The effect of a reduction in required SNR associated with the SCA channel is a dB-for-dB change, which can be substituted into Step 3. Thus, reducing the required SNR on the SCA channel by only 5 dB, because of the more robust digital signal, will restore the SCA coverage to that of the main "entertainment" signal for the FM station.

MODULATION AND ERROR CORRECTION

SCA regulations allow freedom to choose any modulation scheme. In general, the more sophisticated the modulation technique employed the higher the bit rate yield of a given bandwidth, measured as bits per hertz. For example, a QPSK modulation will offer double the channel capacity compared with bipolar or 2-PSK. However, the engineering trade-off is the cost of higher requirements of power radiated, that is, better SNR at the receiver end to obtain the same BER. Nevertheless, keeping the same emitted power level or received SNR, the BER can still be maintained by incorporating appropriate error correction schemes into the system to obtain an ultimate channel capacity that is much better than that available with 2-PSK, although it may not be twice as much.

Out of a number of modulation schemes that exist, such as minimum-shift keying (MSK), multiamplitude MSK (MAMSK), offset PSK, $\pi/4$ shifted PSK, raised cosine PSK, quadrature amplitude modulation (QAM), nonlineraly filtered QAM, Gaussian filtered MSK (GMSK), and so on, one or more may have an optimal set of characteristics to make the best fit for STIC in combination with one or more error correction schemes. In a mobile environment with Raleigh fading and multipath in an urban surrounding, unacceptable negative impact on the received BER would occur without robust signal properties and conditioning. This calls for thorough analytical and, perhaps, simulation studies to analyze a variety of scenarios and arrive at the right choice of modulation and error correction schemes that would provide an acceptable error rate and highest possible data capacity, accompanied by a reasonable cost for user equipment. It is important to also bear in mind the fact that implementation of not all modulation (for example, "codulation" approach or modulation in combination with memory) techniques may be currently available in chip form, but because of the volume of motorists, if the services are found beneficial and marketable, the volume production itself has the potential of offering the user equipment at a much reduced price.

BENEFITS OF THE STIC DESIGN

- It is available quickly for early field trials and initial operational systems.
- It has adequate capacity to handle IVHS growth for the foreseeable future.
- No additional RF spectrum is required (procurement processes for radio frequencies are long and cumbersome).
- It has very low infrastructure capital and OA&M costs and relatively inexpensive user equipment.
- The user of low-end (basic traffic information on congestion, incident, weather, and safety advisory) IVHS services, with relatively inexpensive in-vehicle hardware and without any elaborate data base capability, still has available location information, associated with traffic congestion or incident, with an intersection-level granularity, for display in abbreviated alphanumeric character form straight from the information broadcast source.
- It is easily integrated into long-term "full-up" IVHS architecture. The basic low-cost hardware in the vehicle is common to both the low- and high-end (navigation, route guidance, etc.) versions. This situation enables the user to scale up subscription from a low-end to high-end set of services when the latter is available and the user is ready to make use of it.
- On the other hand, motorists who normally utilize the high-end set of IVHS services in their hometowns may prefer to fall back on the low-end set of IVHS services while traveling to (or on a long-distance trip passing through) some other town temporarily, instead of buying the detailed (CD) data base relevant to the particular city of travel. For the high-end family of IVHS services, one would require additional equipment—for example, processors and data bases, and so on.
- Many other types of services could be subsumed within the STIC system, for example, for transit information and correction information for differential global positioning system. The SCA channel capacity is expected to have enough capacity for quite a few years in the future. Furthermore, if the need for greater capacity arises later, there are comparatively simple ways of updating the STIC system. First, more sophisticated and efficient modulation and coding schemes are already being conceived, developed, standardized, and implemented on chips. Second, one will have the option of arranging for additional FM broadcast stations for transmission of traffic information while incorporating regionalization into the data to be broadcast.
- A major advantage of the STIC system is its flexibility to simultaneously accommodate the users of both the low- and high-end IVHS services; in other words, both types of motorists will be able to coexist. Consequently, there is assurance for soft transition no matter how the user population changes: (a) only low-end user penetration changes, (b) only high-end user penetration changes, or (c) any combination in any proportion of the above two changes.

CONCLUSIONS

STIC can provide a large-bandwidth data communications link between the highway infrastructure and vehicles. This link could be implemented very quickly to support IVHS field operational trials and would integrate well into an end-state IVHS architecture.

REFERENCES

1. J. Chadwick, V. Patel, and G. Beronio. The Relationship Between System Architecture and Radio Frequency Requirements for Intelligent Vehicle Highway Systems (IVHS). *Proc., 1992 Annual Meeting of IVHS AMERICA*, Newport Beach, Calif., 1992.
2. M. P. Ristenbatt. *A Communications Architecture Concept for ATIS.* Communications and Signal Processing Laboratory, Electrical and Computer Science Department, University of Michigan, Ann Arbor, 1991.
3. *Mobile Information Systems.* (J. Walker, ed.) Artech House Publishers, Boston, Mass., Oct. 1990.
4. J. Chadwick and V. Patel. A Communications Architecture Concept for Intelligent Vehicle Highway Systems (IVHS). *Proc., Intelligent Vehicles 1992*, Detroit, Mich., July 1992.
5. P. Davies and G. Klein. Field Trials and Evaluations of the Radio Data System Traffic Message Channel. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991.

Publication of this paper sponsored by Committee on Communications.

Integration of Machine Vision and Adaptive Control in the Fast-Trac Intelligent Vehicle Highway System Program

PANOS G. MICHALOPOULOS, RICHARD D. JACOBSON, CRAIG A. ANDERSON, AND JAMES C. BARBARESSO

Machine vision has been considered one of the most promising technologies for developing and deploying advanced traffic management system and advanced traveler information system applications. Although experimental video detection systems are being developed worldwide, a real-life, practical, and effective application has been elusive. The first and largest application of video detection and its integration with adaptive control in a network of 28 intersections in Oakland County, Michigan, are described. The broader project (called FAST-TRAC), which includes a driver information system, is also summarized. The successful deployment of the state-of-the-art detection/adaptive control system led to the decision to expand it to 93 intersections by June 1993 and ultimately to 800 by 1996.

Advanced traffic management systems (ATMS) is one of the most important areas of the intelligent vehicle highway systems (IVHS) program recently established in the United States. Although many ATMS programs are in the planning stage, the Road Commission for Oakland County (RCOC), Michigan, has successfully deployed the first ATMS in North America that utilizes adaptive, coordinated traffic control of intersections driven exclusively by a new technology, namely video image processing for vehicle detection. The adaptive control system is the Sydney Coordinated Adaptive Control System (SCATS), developed in Sydney, Australia (*1*), and the video vehicle detection system is the AUTOSCOPE (*2*), developed at the University of Minnesota with funding from the Minnesota Department of Transportation, FHWA, and the Center for Transportation Studies at the University of Minnesota. AUTOSCOPE was recently comercialized by Image Sensing Systems and marketed by Econolite Control Products. A third component, an advanced traveler information system (ATIS) called Ali-Scout (*3*), provides an infrared communication link between vehicles and roadside beacons that allows for continuous exchange of traffic and route guidance information, with vehicles also acting as probes to determine travel times. The Ali-Scout was developed by Siemens.

P. G. Michalopoulos, Department of Civil Engineering, University of Minnesota, 500 Pillsbury Dr. S. E., Minneapolis, Minn. 55455. R. D. Jacobson and C. A. Anderson, Image Sensing Systems, Inc., 1350 Energy Lane, Suite 2, St. Paul, Minn. 55108. J. C. Barbaresso, Road Commission for Oakland County, 31001 Lahser Road, Birmingham, Mich. 48010.

The RCOC program responsible for the deployment of these IVHS technologies is called FAST-TRAC (Faster and Safer Travel Through Traffic Routing and Advanced Controls). The program's primary mission is to deploy and evaluate operational ATMS and ATIS technologies that lead to improved mobility and safety on increasingly congested arterial roads and freeways in Oakland County, Michigan. The program has proven to be an excellent example of how public, private, and academic institutions have cooperated in the planning, design, and successful deployment of state-of-the-art traffic technologies. Further background and information on the overall programs basis, objectives, and goals are described elsewhere (*3*).

In this paper the field testing and implementation of the video vehicle detection system in Oakland County and its integration with SCATS are presented. This includes the engineering effort and system performance to meet or exceed the vehicle detection requirements for SCATS intersection control.

FAST-TRAC PROGRAM SUMMARY

The Oakland County IVHS Program, FAST-TRAC, is a large-scale deployment of an integrated ATMS and ATIS. In what follows, the goals of FAST-TRAC are briefly described along with its vision, plans, and status.

The major objective of FAST-TRAC is to demonstrate the effectiveness of an integrated ATMS-ATIS system in improving mobility, reducing energy consumption, enhancing air quality, and reducing traffic accidents. The program also will evaluate and demonstrate requirements for deployment of IVHS throughout North America. Technological constraints and considerations for deploying such systems will be identified and evaluated. Organizational requirements for deploying IVHS at the local, regional, and state levels and across multiple jurisdictions will be defined.

Multiple jurisdictions are becoming involved in FAST-TRAC. Not only will the traffic signal systems of local units of government be integrated, but freeway operations under the control of the Michigan Department of Transportation

(MDOT) will also be incorporated. A link between MDOT's Metropolitan Transportation Center and the Oakland County Traffic Operations Center will be made. Freeway ramp controls, variable message signing, and other freeway operational improvements will be integrated with the surface street ATMS-ATIS to guarantee optimal performance of the overall highway system.

Data requirements and methods of reporting for effective traffic management will be reviewed. The deployment of these technologies will alter staffing requirements and the expertise that staff will be required to have. Finally, the affordability of these systems, especially for local implementation, will be evaluated in a later phase, as maintenance life-cycle costs are accumulated.

The FAST-TRAC partners have developed a program that satisfies many of the operational test objectives spelled out in the IVHS America Strategic Plan. In the ATIS area, commercial, public service, public transit, and private vehicles will be equipped to allow the assessment of human factors, considerations, and productivity gains. These vehicles will also act as traffic probes, providing real-time information for incident management and dynamic route guidance.

It is a goal of FAST-TRAC to provide a focal point for the collection and dissemination of traveler information on an areawide basis. A traffic operations center (TOC), currently under construction, will provide this focal point and will be linked to the Metropolitan Transportation Center. The ATMS is designed to enhance political and institutional cooperation for effective, areawide traffic management.

FAST-TRAC includes a third-generation traffic management system and uses video image processing for traffic detection. Other IVHS technologies, such as advanced communication technologies, will be used in FAST-TRAC. Alternative communication technologies will be tested to link the TOC to intersection hardware and to make the vehicle-to-roadside communication link.

FAST-TRAC has been made possible through funding from a variety of sources. Initial funding for the first phase of the project was contributed by the Oakland County Road Commission, direct appropriation from Congress, Siemens, and MDOT.

The program is broken down into four phases spanning a period of 5 years. Phase I, which was completed in June 1992, was a pilot phase in which SCATS and AUTOSCOPE were integrated, tested, and deployed at 28 intersections, 23 of which are fully instrumented with AUTOSCOPE and the remainder with loops. Phase II currently consists of 95 intersections in the city of Troy, Michigan, that includes all of southeast Oakland County. The third phase of the program will involve the deployment of another 100 intersections in the burgeoning area surrounding the Oakland Technology Park and the Pontiac Silverdome. Ultimately, all of the urbanized portion of Oakland County will come under SCATS and AUTOSCOPE operations (800 intersections). Some rural applications also will be tested.

A TOC will be established for the combined ATMS-ATIS. The TOC will house the central computers and communications equipment for managing and monitoring the systems. It will also include meeting and training facilities. The TOC is currently being built and is described later in the paper.

AUTOSCOPE VIDEO DETECTION SYSTEM

Only a short description of the system is included here. The interested reader can turn to previous papers (*2*, *4*) for more detailed description of AUTOSCOPE. Briefly, the AUTOSCOPE system can detect traffic in multiple locations of areas within the field of camera's view. These locations are specified by the user in a matter of minutes using interactive graphics and can be changed as often as desired. This flexible detection placement is achieved by placing detection lines along or across the roadway lanes on a TV monitor displaying the traffic scene using a mouse. Therefore, these detection lines are not physically placed in the pavement but only on the TV monitor and can be removed after initial placement of the detection lines. Every time a car crosses these lines, a detection signal (presence or passage) is generated by the device. This signal is similar to that produced by loop detectors. Thus, the system emulates loops and can easily replace them in existing installations. However, the advantage is that, in addition to the wireless detection, a single camera can replace many loops, thus providing a true wide-area detection and becoming cost-effective.

SYSTEM OVERVIEW

As mentioned earlier, the current Phase I installation consists of 28 intersections. This small-scale network is the pilot testing of the integrated AUTOSCOPE and SCATS control systems. Of the 28 intersections instrumented, 23 exclusively use AUTOSCOPE for vehicle detection. The remaining five minor intersections require minimal detection coverage and have sufficiently stable pavement to install loops.

SCATS (*1*) is a complete traffic management system that provides adaptive, real-time traffic signal control, coordinated on a network basis. The system has been developed over a period of almost 20 years to control traffic signals in and around Sydney, Australia. Oakland County, Michigan, is the first North American installation of SCATS. The control approach is demand responsive by dynamically adjusting cycle lengths, splits, and offsets for coordination. Cycle lengths are varied (anywhere from 20 to 190 sec) on the basis of overall demand; split adjustments are based on demand between competing directions; and offsets coordinating adjacent intersections are varied with traffic demand to minimize the number of stops within the network. The basis for determining demand is the degree of saturation (DS). The DS is the ratio of effectively used green time to total available green time. The control strategy is designed to minimize "wasted" green time throughout a system to suit the prevailing average traffic conditions.

The DS calculation has some unique vehicle detection requirements. When SCATS is implemented through loops, the vehicle detector size is required to be 4 to 4.5 m long to compute consistent DS results for all vehicle speeds and congestion levels. The detectors must also be placed 1 m back from the stopline for adequate coverage of stopped vehicles. The detectors are almost exclusively placed at the stopline and are used for two purposes: (a) tactical detection to enable a vehicle movement phase such as a left turn and (b) strategic detection to dynamically adjust cycle, split, and offset times. Strategic detector information can be used to control intersections where they are placed as well as intersections downstream of their placement, to measure approaching traffic demand. Detectors can be used for both tactical and strategic purposes, depending on the lane of travel and the complexity of the downstream intersection that the lane feeds.

A typical placement of SCATS/AUTOSCOPE detectors in Oakland County is shown in Figure 1. The figure shows where detectors are placed at intersections upstream of the controlled intersection for strategic inputs for adjusting cycle length, splits, and offset times; it also shows the camera placements for AUTOSCOPE detection.

The size and placement of the SCATS-compatible detectors had serious maintenance and reliability concerns if loop technology were to be used. The road surfaces in Oakland County are typical of those of northern U.S. cities with an assortment of various-sized potholes and combinations of concrete, pavement, and patches. The 4.5-m length implies that loops must typically span mutiple cement pads and survive seasonal shifting of the road surface. Road weight restrictions are few, and it is common to see vehicles with as many as 40 wheels navigating the roads with enormous loads. The average temperature in the region during fall, winter, and spring is near freezing (ice storms are frequent near the Great Lakes). The freezing and thawing cycles destroy the road surfaces very quickly, especially when raw salt is poured from local salt mines onto the road. The survivability of long loops in this environment was in serious doubt; that is, it was questionable whether they would be able to last longer than 1 year at a time.

The aforementioned problems, along with the following considerations, led to the decision to use AUTOSCOPE in the FAST-TRAC project:

1. The detectors are nondestructive to the road surface, with no lane closures to install or maintain equipment.
2. The multiple detection capability offered by AUTOSCOPE can be used for queue length detection that will be used to develop the next generation of SCATS; when this occurs, no new detector installation will be required.
3. The AUTOSCOPE and peripheral equipment can be installed any time of the year. Most of the AUTOSCOPE installation for this initial project occurred from February through May.
4. AUTOSCOPE can be used in the future for deriving stops, delays, energy consumption, and pollution levels for continuous monitoring of SCATS performance.
5. The detectors can meet the size and placement requirements for SCATS. A study was conducted to define and test AUTOSCOPE detectors that emulate the area coverage of a 4.5-m loop placed 1 m behind the stopline; the performance results are presented later.
6. The detectors can be positioned over any road surface or combinations of road surfaces with no loss of performance or consistency.

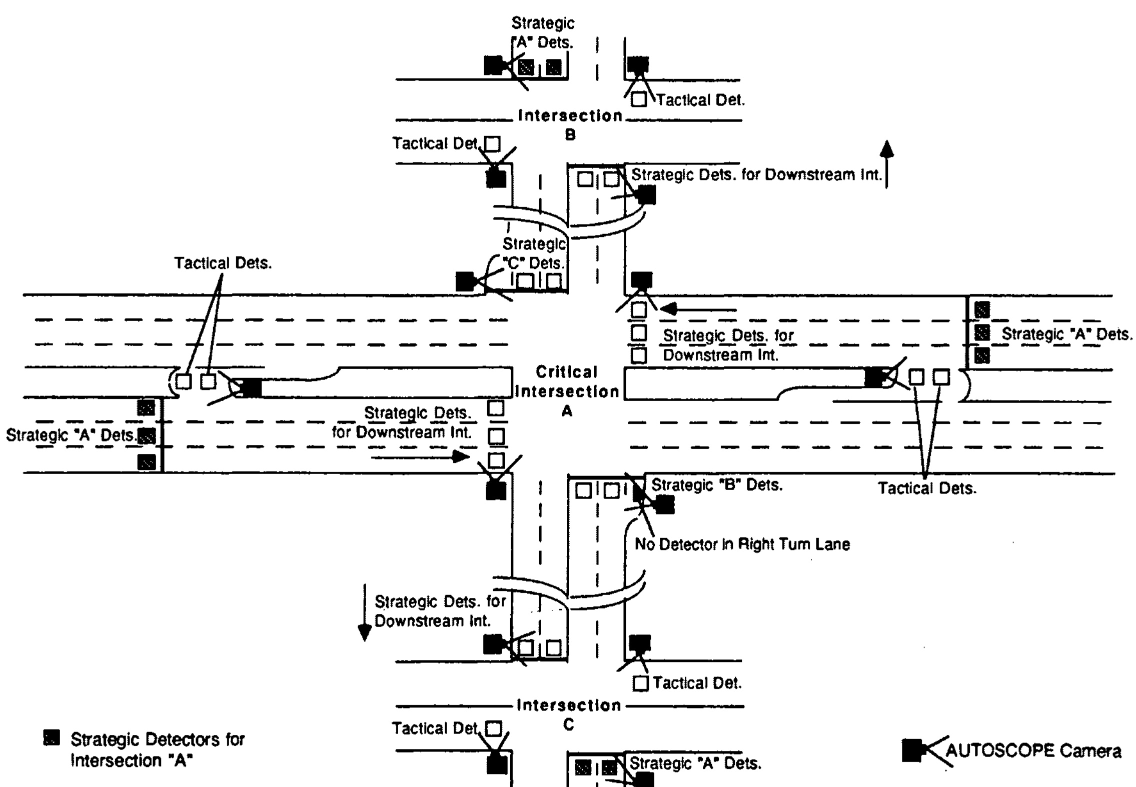


FIGURE 1 Typical SCATS-controlled intersection configuration.

7. The detection performance of AUTOSCOPE has been documented to perform under all weather, lighting, and range conditions that do not obscure the camera's field of view.

Figure 2 shows how the FAST-TRAC system is configured. AUTOSCOPE cameras are placed at designated locations to acquire the desired road coverage for SCATS tactical and strategic detectors. The AUTOSCOPE detectors are tied into the SCATS local controller using a NEMA TS/1 contact closure interface. Several Ali-Scout roadside beacons are at several intersection approaches to exchange information with Ali-Scout–equipped vehicles that pass through the intersection and passed on to the Ali-Scout central system. The SCATS local controllers are tied directly to the SCATS regional computer via dedicated phone lines (and modem). The local SCATS controllers execute timing plans defined by the regional computer, thus allowing synchronization and coordination of local controllers in a region. The SCATS central management system provides coordination between regions and a common data base for the network. AUTOSCOPE vehicle detections, traffic parameters, and statistics are managed by the SCATS. The SCATS regional computer also has a user interface to monitor and manage the SCATS controlled intersections. It currently acts as the primary input for any command and control activities. Phases III and IV of the FAST-TRAC program will integrate the SCATS hardware with the Ali-Scout and video surveillance for a unified command and control display console. All command, control, and response functions will be activated via a geographical information system user interface. The command and control system will be interfaced with other traffic information centers, highway patrol, and MDOT to exchange information, manage incidents, and coordinate the management of freeways and arterials via ramp control, changeable message signs, and vehicle communication.

PRELIMINARY ENGINEERING

Because the SCATS and AUTOSCOPE technologies were new to the engineers and technicians of Oakland County, RCOC decided to expand the funding for the preliminary engineering so that the systems could be properly deployed. Preliminary engineering for the SCATS controller was relatively straightforward because it is a very stable product that has been in service for many years. The majority of engineering effort for the SCATS was user training and defining control designs for the various intersection test sites.

The major engineering issue to resolve was using AUTOSCOPE as the vehicle detection sensor. The AUTOSCOPE was a proven prototype (*2*) but was not field ready when the project began in August 1991. The AUTOSCOPE needed to be environmentally hardened to meet industrial (NEMA) specifications for temperature, shock, radio frequency, form factor, and so forth. Cost-effectiveness would also be gained by providing up to four cameras of video input into a single AUTOSCOPE. The AUTOSCOPE product development was not funded directly by RCOC but offered an excellent set of system requirements that the AUTOSCOPE had to satisfy. In view of this, development was funded privately by Image Sensing Systems and Econolite Control Products. In this manner, the private sector was brought in to complete a product for which the basic research and development were performed in a university environment with government funding.

There were several preliminary engineering tasks that had to be performed to weatherize and develop AUTOSCOPE. The first was to plan and coordinate the trade-off and analysis studies, hardware selection and procurement, hardware installation, site detector layout programming, detector data file management, and so on. The second was to define, test, and evaluate the performance of the AUTOSCOPE to successfully emulate the detection requirements of the loop-based

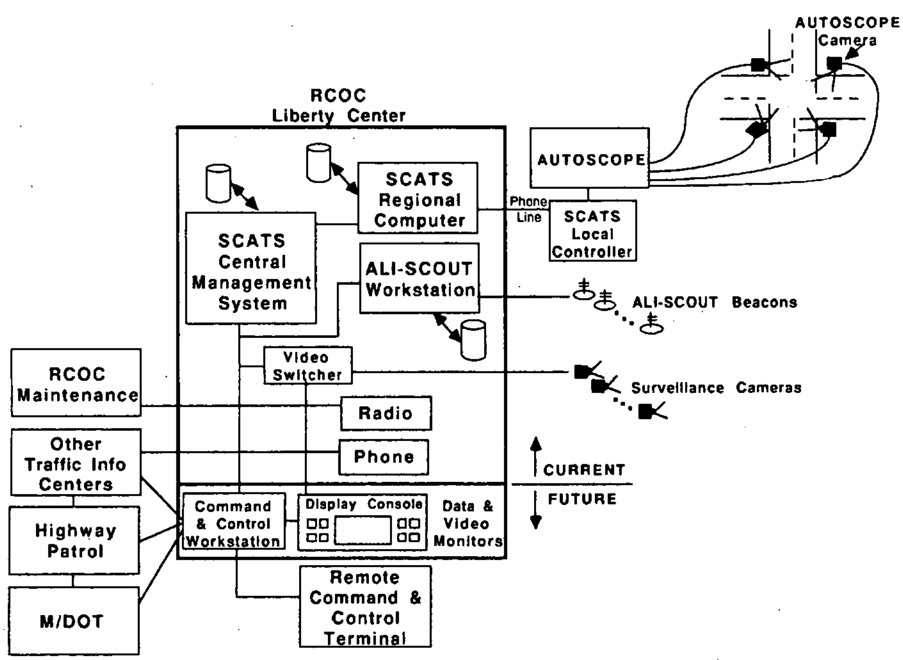


FIGURE 2 RCOC command and control center.

systems deployed in Sydney, Australia. There were two requirements for meeting this objective. The first was that the detector size and placement provide consistent performance so that the SCATS degree of saturation control signal is reliable. The second requirement was that the absolute detection of vehicles at stoplines be robust to safely and efficiently control turning and through movements.

The third task in the preliminary engineering was to perform site engineering of detector placement for each intersection approach. This included definition of the infrastructure to support the AUTOSCOPE and peripherals (poles, camera, optics, cabinets, cables, power, etc.) for the desired detector placements. Details are presented later.

The last major task during the preliminary engineering was to develop the interfaces between the SCATS and AUTOSCOPE so that they are electrically compatible, provide the correct contact closer characteristics during system operation (and nonoperation), and are reliable. A loop-compatible output module for the AUTOSCOPE that meets NEMA TS/1 specifications was developed, tested, and deployed to meet this requirement.

SITE ENGINEERING

Site engineering for the AUTOSCOPE is very different from what traffic engineers have experienced before. However, the engineering concepts are easy to understand and practice. One of the most powerful impacts of the technology is the human visualization and verification (instant feedback) that this video detection system provides. The order of engineering tasks to perform for implementation generally follows the steps of the FAST-TRAC deployment. The first is to get a thorough understanding of the detection requirements of the application to be performed, that is, where detectors are needed and what type (volume, occupancy, speed, gap times). For the SCATS installation, it was necessary for the SCATS engineers and RCOC to define their desired control strategies before any AUTOSCOPE site engineering could begin. Next, one should define a detector placement configuration strategy that determines the camera's expected field of view. Some detectors and detector information are typically more important than others in the configuration. It is important to understand which detector functionality may have to suffer because of geometric road considerations, such as safe placement and easy access to poles for maintenance, absolute stopline detection to enable safe turning movements, and so on. Knowing the priority of detection needs makes it easier to choose optimal camera locations. The third step is to take advantage of existing infrastructure where and whenever possible; that is, use existing poles, span wires, underground conduit, cabinets, and so on. When new equipment must be installed, providing for public safety and easy maintenance access is necessary.

The engineer should also update site drawings of existing sites to reflect "as-built" dimensions. The following dimensions are of most interest: the accurate placement of stoplines from the nearest corner, existing poles in the vicinity and dimensions to the road, and height on the poles where cameras can be safely mounted without interfering with power or phone lines. There were basically four types of intersection configurations that existed in Oakland County. The first is a boulevard intersection with three to four lanes on the boulevard and two to three lanes on the crossroad with no left turns on any approach. The second is a five-lane intersection with two through lanes per approach on the major road with a fifth lane used for protected left turns. The minor road typically would have one to two through lanes per approach with protected or permissive left-turn lanes. The third intersection configuration is a minor intersection with two through lanes per approach on the major road with permissive left turns and one to two lanes on the minor road. The phase for the minor road was typically enabled by nonlocking detectors. Finally, the last configuration involves boulevard crossovers to eliminate protected left turns on boulevards. The crossover phases are enabled by nonlocking detection. The area coverage was typically very large. The stopline was 9 to 12 m wide to accommodate turning trucks. Vehicles could approach the stopline anywhere along the stopline. An example of these four intersection types is shown in Figure 1, which shows where SCATS detectors are typically placed and where the detector information is used.

RCOC was willing to route video and power lines across span wires (above ground) at the majority of intersections so it was not necessary to install underground conduit or to check if there was existing capacity for the video and power cables.

Another step during the site engineering is to define a camera location and height that minimize the field of view obstructions (occlusions) on the basis of the camera's location and perspective. Increasing the camera height reduces the occlusion effects. Drawing some simple diagrams of camera height, number of lanes of coverage, pole distance from the road, and expected vehicle heights from occluding lanes typically is used to define a required camera height. The camera placement should also minimize reflections from leading headlights and adjacent roads at night. Leading headlight reflections can be eliminated by observing vehicles moving away; this, however, results in unrealistic and expensive "extra" pole placements at intersections. As an alternative, headlight reflections can be minimized by placing detectors in closer downlane proximity to the pole, thus reducing the amount of reflected light that enters the camera. In addition to this, enabling processing algorithms in the AUTOSCOPE were developed that significantly reduced the effects of leading headlights. Finally, the majority of poles and cameras placed in Oakland County were chosen to be adjacent to the stopline at the same approach. This location optimizes the field of view for high-accuracy stopline detections (volume and occupancy) and minimizes leading headlight reflections that can result in false detections. Luminaires at the intersections also provide excellent occupancy results at night.

AUTOSCOPE PERFORMANCE EVALUATION

The performance of the AUTOSCOPE has been evaluated previously on freeways (*2,4,5*) and most recently at an intersection (*5*). Each evaluation included long-term data comparisons of AUTOSCOPE detection versus loops and manual ground truthing. Comparisons of the loops and AUTOSCOPE with manual ground truth show comparable performance and accuracy levels in excess of 95 percent for volume, speed, and occupancy. Although AUTOSCOPE can be used in conjunction with any control strategy, the special requirements of the SCATS detectors necessitated additional testing and calibrations to ensure full compatibility. The results of this testing are presented here.

To verify that it could meet or exceed the unique vehicle detection requirements of the SCATS and intersection control applications for the Oakland County installation, AUTOSCOPE was further tested and modified to determine its detection performance and adapt it to the following problem and artifact conditions.

1. Intersections impose stiffer safety requirements to absolutely detect vehicles at stoplines and to not remain stuck on when no vehicles are present (which leads to overall system inefficiency).
2. SCATS has requirements of high-accuracy volume and occupancy to compute reliably the SCATS degree of saturation control parameter.
3. Detection must be robust to work under all weather and lighting conditions (day or night). Shadows from vehicles in adjacent lanes pose the biggest problems here. Almost every video sequence analyzed had some form of weather artifact, such as snow, rain, sleet, ice, fog, high winds, and so on. There were some concerns of obtaining accurate occupancy at night. In all but two approaches at one intersection, there was sufficient illumination from street lights, commercial or retail business lights, and closely following vehicles to obtain very good occupancy at the stopline for the 23 intersections in Oakland County.
4. Downlane and crosslane geometric occlusion effects were minimized by detector placement at the stopline. Only five poles were added or moved on the entire project of the 62 cameras that currently are in operation. The existing pole infrastructure was used cost-effectively to retrofit existing intersections.

Several improvements were made to the detection algorithms to meet the above detection requirements. A new presence detector was defined that is very effective at holding and maintaining a vehicle detection whenever a vehicle passes through or stops under the detector for long periods of time, thus providing true presence. The presence detector is typically placed downlane at the stopline to obtain good coverage of the roadway for vehicles that stop short of the stopline. A further improvement to the AUTOSCOPE presence detector that was instrumental in fulfilling deployment detection requirements was the ability to distinguish which direction the vehicle enters the detector or is stopped under the detector. This capability immediately eliminated all false actuations due to "wrong-way" detections. Such detections were caused by strong shadows entering the detector from adjacent lanes, turning vehicles from other approaches taking shortcuts across empty left-turn lanes, and any light reflections (day or night) that enter the detector from the wrong way (such as an exit of a convenience store across the road from the detectors that shines light across the detectors).

The AUTOSCOPE has another type of presence detector called a count detector that is used for counting closely following vehicles at range. The detector is typically oriented across the road (crosslane) to separate closely following vehicles at range. The count detector is better suited for freeway applications where there is typically some vehicle motion, even in the heaviest of congestion. The count detector has not been required by SCATS so far.

The performance of AUTOSCOPE was analyzed with the directional presence detectors in through and left-turn lanes. Some of the video sequences analyzed were taken from non-optimal camera positions so that performance degradation could be determined. The AUTOSCOPE detector presence was compared with ground truth presence as determined by visual observation of vehicles within the detection zone on a frame-by-frame basis. Frame rates are on 33-msec boundaries, and it was estimated that the human observer could distinguish the entry and exit time of the vehicles to within one to two frame times. This was accomplished by slow-motion videotape replay methods.

The performance of the AUTOSCOPE directional presence detectors placed in through lanes at several intersections using the SCATS size and placement specification of 4.5 m long and 1 m behind the stopline is shown in Table 1. All performance measures are given in percentages compared with the manually determined ground truth data. The detection accuracy is the percentage of ground truth vehicles that were detected. The false detection rate is the sum of the number of double detections of the same vehicle plus the number of false detections with no vehicle present, given as a percentage of the number of actual vehicles. The volume

TABLE 1 AUTOSCOPE Direction Detector Through Lane Performance Compared with Manual Ground Truth

TEST ARTIFACTS	DET. ACCURACY %	FALSE DET. RATE %	VOLUME ERROR %	OCCU-PANCY ERROR %	NUMBER GROUND TRUTH VEHICLES
1. High Winds, Shadows	98.1	3.0	1.1	-2.5	264
2. Overcast, Snow, Windy	99.6	4.2	3.8	-2.7	262
3. Overcast, Wet Rd, Windy*	96.1	2.0	-2.0	-6.1	51
4. Day/Night Transition	97.1	7.4	3.5	4.0	312
5. Night	98.1	6.1	4.2	7.5	407
6. Partly Cloudy, Lt Wind	96.2	4.1	0.3	-3.7	290

*Not an actual camera location; viewed head-on at an exit ramp from across the intersection road.

error is the difference between the total number of detections and the number of vehicles, given as a percentage of the number of vehicles. Finally, the occupancy error is the difference between the number of frames where the detector was on and the number of ground truth frames where a vehicle was judged to occupy the detection zone, given as a percentage of the number of ground truth frames on which the detection zone was occupied. A negative sign of volume and occupancy error implies that the AUTOSCOPE value is lower than manual ground truth.

The performance of the AUTOSCOPE directional presence detectors placed in left-turn lanes was also evaluated. The number of cars missed while the detector was on and off was counted. A miss while on means that the detector was already on when the vehicle entered the detection zone, as it would be, for example, for a closely following vehicle; therefore, the number of separate actuations was not incremented for that vehicle. A miss while off means that the detector did not detect the vehicle while it was present. A miss while off is of crucial importance for left-turning lanes because errors cannot be tolerated for safety reasons. The tests consisted of several 30- to 45-min videotape sequences for each artifact. Each intersection tested was running at approximately 90-sec cycle lengths. The seemingly large number of misses with the detector on is a desired behavior of the detector for SCATS. The detector was sized at 3.5 to 4.5 m to bridge the gap between two closely following vehicles traveling at less than 13 km/hr (8 mph). It should be evident that, with detectors that long, it is practically impossible to separate cars when they are closely spaced. Thus, the number of misses with the detector on was a desirable feature for SCATS.

The only two misses confirmed so far with the detector off are considered critical failures of the detector. Close examination of the two cases indicated that the vehicles partially passed through the detector by shortcutting the left turn. As a result of this analysis, two AUTOSCOPE directional presence detectors were placed side-by-side in every left-turn lane approach. So far the two-detector configuration has not failed to enable the left-turn movement. The only reported cases of failure to enable the left turn were because drivers were not pulling up under the 4.5-m detectors placed 1 m behind the stopline. Additional detectors have since been added further behind the stopline to eliminate this problem.

The true test metric for AUTOSCOPE to meet is an accurate extraction of the SCATS DS. The DS (ignoring premature-phase termination by gap time out) is defined as

$$\mathrm{DS} = \frac{Gt - \{T - [h * (N + 1)]\}}{Gt}$$

where

Gt = signal phase green time (sec),
T = total off time of the detector during Gt (sec),
h = minimum time headway, and
N = number of vehicles during the signal green time.

The minimum time headway is the one at which two closely following vehicles would travel at peak saturation. This parameter is recomputed every day on the basis of the largest demand for the day. This value is very sensitive to road geometrics and driver behavior. A typical value is 0.8 sec. Finally, this vehicle count is incremented by 1 to account for the detector size, which is designed to bridge the gap between the first two vehicles when the queue discharges.

The video sequences of Table 1 were analyzed by computing the DS control parameter. Figures 3 through 5 show the magnitude and difference between the manual ground truth and AUTOSCOPE in computing the SCATS DS under only a few adverse conditions. Similar figures for other artifacts, such as night, occlusion, partially cloudy weather, lightning, heavy rain, and so on, are not shown because of space limitations. The magnitude of DS in the figures indicates the congestion levels the detectors were tested under and how closely they track the manual ground truth. The major deviations of the AUTOSCOPE from ground truth were because of incorrect vehicle counts (mostly extra false detections and double counts). However, these deviations are negligible, and even with potentially high false detection rates the difference between the AUTOSCOPE DS and the ground truth DS does not have a significant effect on the intersection efficiency.

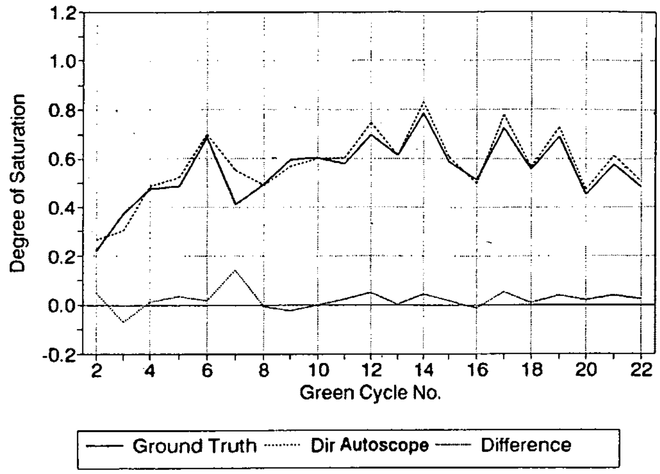


FIGURE 3 Comparison between measured and AUTOSCOPE-derived DS under high wind, strong shadows, and congestion.

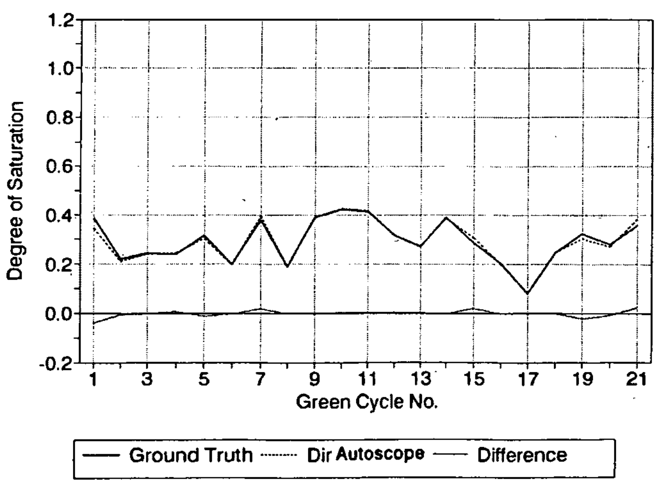


FIGURE 4 Comparison between measured and AUTOSCOPE-derived DS under overcast, snow, wet road, and windy conditions.

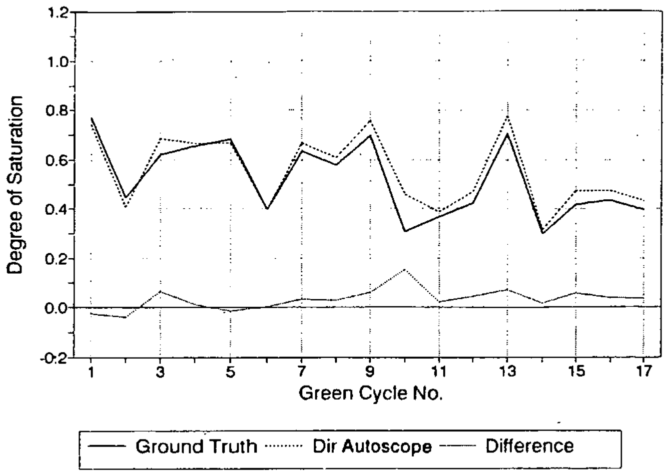


FIGURE 5 Comparison between measured and AUTOSCOPE-derived DS during day-to-night transition.

The results presented here along with the detailed tests were presented to the SCATS developers, who, following additional manual observations of their own, were satisfied that AUTOSCOPE was performing well. Following this verification AUTOSCOPE and SCATS were fully integrated and deployed.

FIELD DEPLOYMENT ISSUES

The installation of the AUTOSCOPE equipment went very smoothly. The biggest problem was in obtaining video connectors to work correctly with the video cable; this problem led to delays in cable installation and replacement of low-quality connectors with the correct connectors for the cable. The cable selected was a new product line because connector manufacturers did not have any connectors that fit the cable properly. RCOC retained the AUTOSCOPE developer under a continuing engineering contract to train installation technicians and engineers, quickly resolve installation problems and issues, define optimal detector placements for the best road coverage for each intersection, assist in programming the AUTOSCOPE detectors, and troubleshoot interface problems with the SCATS controller. Some other minor problems with installing the first AUTOSCOPE production units were quickly overcome. The installation was started in February 1992 and finished in May 1992. The official commencement of the entire system was June 1992.

CONCLUSIONS

The first and largest worldwide installation of video-based vehicle detection for adaptive intersection control was successfully deployed and is operational today. The system has been running continuously since May 1992. Because the integration and field deployment of AUTOSCOPE with adaptive control appears to be an early winner in the area of IVHS, the system is being expanded to 94 intersections and will eventually grow to 800.

Considerable logistical and technical problems had to be overcome. These were resolved through close cooperation of the participants and competent project management. RCOC took considerable, but calculated, risks in attempting to deploy machine vision on such a large scale for the first time rather than seeking the safety of proven systems. The risk seems to be paying off since development of the next generation of adaptive control is now feasible, and plans are currently under way to achieve this control on the basis of the AUTOSCOPE capabilities that have not yet been fully used. RCOC was also prudent to receive thorough training and expert assistance to correctly engineer the system components and installation sites for the deployment of the new integrated technology. The project clearly benefited by having enthusiastic and high quality human resources, including managers, engineers, and technicians to quickly and efficiently make changes and modifications as needed. The entire program is an excellent example of public, private, and academic cooperation to meet the primary program objective of improving urban traffic mobility and safety.

A word of caution on the AUTOSCOPE implementation and field deployment is in order. Specifically, the device is not simply a replacement of loops that will continue to serve their intended purpose for some time; instead AUTOSCOPE should be viewed as a wide-area detection device that can obtain valuable information and extract traffic parameters and measures of effectiveness (delays, stops, queue lengths, energy consumption, etc.) that are hard, labor intensive, time-consuming, and expensive to obtain. Such capabilities should lead to more effective ATMS and ATIS applications and deployment.

REFERENCES

1. P. R. Lowrie, SCATS Sydney Coordinated Adaptive Traffic System: A Traffic Responsive Method for Controlling Urban Traffic. *Road and Traffic Authority of N.S.W.*, Sydney, Australia, 1990.
2. P. G. Michalopoulos. Vehicle Detection Through Video Image Processing: The AUTOSCOPE System. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, 1991, pp. 21–29.
3. J. Grubba, J. Haugan, and R. Knockeart. Implementing the FAST-TRAC ATMS/ATIS Demonstration Program. *Proc., 2nd International Conference on Vehicle Navigation and Information Systems (VNIS)*, Detroit, Mich. 1991.
4. P. Michalopoulos, B. Wolf, and R. Benke. Testing and Field Implementation of the Minnesota Video Detection System. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 176–184.
5. P. G. Michalopoulos and R. Jacobson. Implementation of Image Processing in Traffic Detection, Surveillance and Control. *Proc., International Conference on AI Application Transportation Engineering*. San Bonaventura, Calif., 1992, pp. 311–325.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Software for Advanced Traffic Controllers

DARCY BULLOCK AND CHRIS HENDRICKSON

A systematic approach to traffic engineering software development could provide significant advantages with regard to software capability, flexibility, and maintenance. Improved traffic controllers will likely be essential for many proposed intelligent vehicle highway system applications. A computable language called TCBLKS (Traffic Control Blocks) that could provide the foundation for constructing real-time traffic engineering software is introduced. This computable language is designed to be configured by a graphical user interface that does not require extensive software engineering training to use yet provides more flexibility and capability than is possible by simply changing program parameters. The model is based on the function block metaphor commonly used for constructing robust and efficient real-time industrial control systems. Adapting this model to the transportation sector permits traffic control applications to be programmed by (a) selecting preprogrammed function blocks from a standard library, (b) configuring block parameters, and (c) connecting blocks to other blocks in the strategy. The software model was implemented in C on an advanced traffic controller platform and demonstrated in real time for applications such as signalized intersection control and ramp metering. In addition, this same software model was used to control the ramp meters along a segment of westbound Highway 50 during a demonstration in Sacramento, California, in November 1992.

Twenty years ago, traffic controllers underwent a technical revolution in the switch from electromechanical systems to solid-state microprocessor systems. With the computing technology available 2 decades ago, the most cost-effective approach for software development was to construct specialized, embedded systems tailored to the traffic control industry. Traffic control logic was programmed using assembly language programs that could read and write bits associated with external sensors and actuators. Initially, these microprocessor-based controllers did little more than their mechanical predecessors. Over time, transportation engineers realized that more and more features could be implemented on solid-state controllers, and they upgraded their software accordingly.

Today there is another turning point in traffic control technology. The advances in microprocessor technology over the past decade have dropped the average cost of computing by roughly an order of magnitude every 5 years (*1*). In contrast, the average cost of a Caltrans Type 170 controller has decreased only 15 percent since 1987 and 27 percent since 1982 (California Department of Transportation, personal communication, July 1992). Off-the-shelf, field-hardened, and affordable equipment is available that rivals the computing power of mainframes from 20 years ago. If computing costs continue to decline in this manner, it will no longer be cost-effective for proprietary transportation computers to compete with mass-produced industrial hardware. Migration to these more powerful computers will allow traffic engineers to make a fundamental change in software development practices. Memory capacity and processor limitations will not impose significant constraints on applications. Instead, traffic engineers can focus on developing an efficient architecture for building systems that are more effective, easier to install, and easier to maintain.

This change should have numerous benefits. Since the inception of the microprocessor-based traffic controller, the software engineering effort devoted to constructing traffic control software has been less than ideal (*2,3*). In general, the state of the practice with current microprocessor software is to write software in assembly language without an operating system, permanently install the software on a chip (i.e., burn it into a ROM), and empirically test programs to see if they work. Provided a suitable programming model can be developed, it is now possible to engineer software for greater capability, flexibility, and usefulness. However, no substantial model has been proposed. Such models are crucial for the evolution of an engineering discipline from a solely craft-based practice (*4*).

Selecting an appropriate software model is particularly important because the development of transportation control systems is multidisciplinary, requiring the interaction of transportation engineers, electrical engineers, software engineers, and government officials. In the past, the coupling between traffic engineering concepts and field implementation has been weak. This paper presents a software model that is designed to address professional communication gaps and the need for more capable and maintainable software. It is based on the "function block metaphor" that is widely used in industrial control systems. This model provides the capability for non-programmers to develop intuitive control software by drawing graphic diagrams on a computer screen and filling in menus. This model is based on a formal, real-time scheduling algorithm that allows the correctness and feasibility of strategies to be formally verified. It has been so successful in the industrial sector that many companies have imposed rules restricting development of custom software and require application developers to use "canned software" applications consistent with the function block model.

The following section introduces as background the rationale and characteristics of an advanced traffic controller hardware platform. The subsequent sections describe the model of traffic control software developed in this study. A final section discusses the status of the proposed software model.

D. Bullock, Department of Civil Engineering, Louisiana State University, Baton Rouge, La. 70803. C. Hendrickson, Carnegie Institute of Technology and Department of Civil Engineering, Carnegie Mellon University, Pittsburgh, Pa. 15213.

ADVANCED TRAFFIC CONTROLLER HARDWARE

Most of the current generation of traffic controllers used in the United States are based on two different families of controllers. This section summarizes these competing efforts and describes some hardware modernization efforts under way.

One family of controllers, referred to as NEMA units, are built with connectors that conform to standard mechanical and electrical connectors. The philosophy of this standard is that manufacturers will compete on the basis of the hardware and software they provide inside the controller. In theory, an agency can migrate to another manufacturer's controller by unplugging the old one and plugging in the new one using standard connectors. Because of additional proprietary sockets added to the NEMA TS1 units and nonstandard communication protocols, this interchangeability is not realized in practice. The 1988 NEMA TS1 standard recently has been updated (NEMA TS2 Type 1 and NEMA TS2 Type 2) to address deficiencies of the NEMA TS1 standard and incorporate an alphanumeric display for interaction with the controller. Because the software on all NEMA controllers remains proprietary and cannot be ported by the customer, engineers and technicians must still learn new software to reconstruct timing and phasing plans on new NEMA controllers.

A second family of controllers, referred to as the Caltrans Type 170 controllers, is built to provide both standard connectors and portable software. The philosophy of this standard is to develop a very precise specification for a traffic control microcomputer. Manufacturers are selected periodically on the basis of competitive bidding. This standard has been successful for the past 20 years. Minor modifications have been introduced over time, including a second serial port, additional memory, and different ROM sizes, but the essential features are unchanged. The distinguishing feature of the 170 controller remains the program module—an insertable card with a ROM that stores the traffic control program. This module can be removed from one manufacturer's 170 controller and inserted into another manufacturer's controller, and the software will run without modification. This decoupling of the hardware and software procurement is very desirable, particularly when equipment purchases are going to be staged over several years. Instead of relying on embedded user interfaces, as in the NEMA controllers, the 170s are typically configured by connecting a small computer such as a personal computer to a serial port and downloading the strategy. Alternatively, binary configurations can be keyed in on a hexadecimal keypad. A modernization of the Type 170 has been undertaken by New York state and is called the Type 179. This controller provides more powerful computing and employs a real-time operating system. However, it has not been widely adopted. As a result, it has been difficult to develop a pool of competitive vendors.

In view of the microprocessor and software engineering developments in the past 2 decades, these standards are beginning to age (*3*). First, the software is written entirely in assembly language. The complex nature of assembly language development precludes all but the largest cities and states from maintaining a software staff for making software configuration changes other than changing parameters in a given configuration. Second, no operating system is employed (except for some 179 software). Routine chores, such as task scheduling, memory management, and semaphores, must be recoded. The "home-grown" executives that have evolved preclude sharing of new control strategies. Third, the hardware constraints (slow processors and limited memory) can only be addressed by a revised standard that would require rewriting large quantities of assembly language applications. Finally, it is unclear how much longer it will be cost-effective for the transportation community to manufacture specialized computers.

These shortcomings and requirements for improved software development tools, faster processors, expandable I/O, and more memory have led the California Department of Transportation to investigate the use of modular industrial computers for applications ill suited to the Caltrans 170s (*5*). This proposed platform, called advanced traffic controller (ATC), is based on a 3U VME bus, a 680x0 processor, and an OS-9 operating system. This computer is used extensively in the military and commercial sectors and provides an economical, off-the-shelf hardware platform for the ATC. Although a rich set of development tools, including operating systems, compilers, and debuggers, are available for this platform, the low-level nature of the tools renders them inappropriate for everyday use by traffic engineers. This is analogous to a desktop computer that has only a language compiler. For a desktop computer to be truly useful to an engineer, application software such as a spreadsheet or CAD package must be available. Because of this ATC software void, a general-purpose application program (software model) is required to enable traffic engineers to develop real-time traffic control strategies.

COMPUTABLE LANGUAGE FOR TRAFFIC ENGINEERING

The motivation for developing a computable language is to provide a high-level configuration tool that does not require extensive software engineering training to use, yet provides more flexibility than just changing program parameters. The model underlying the language proposed in this paper is based on function block programming in which the function blocks specialized to traffic engineering are graphically assembled and downloaded to a field controller. In the function block programming paradigm, a user develops applications by selecting and connecting predefined software modules called blocks. The blocks represent parameterized programs prepared in a uniform manner, which permits them to be interconnected with other blocks. Connections between blocks serve as communication links for particular variables such as detector states, approach volumes, or phase timing. Selection of blocks may require definition of parameters, such as execution frequency, minimum and maximum green extensions, and filter times. Function block programming lends itself readily to graphical displays in which blocks are represented pictorially as a box with a title, indicating the program associated with a particular block, and a name, providing a symbolic means of referring to elements of a specific block. Figure 1 shows an example of a simplified semiactuated signal with presence detectors on the east and west approaches. Con-

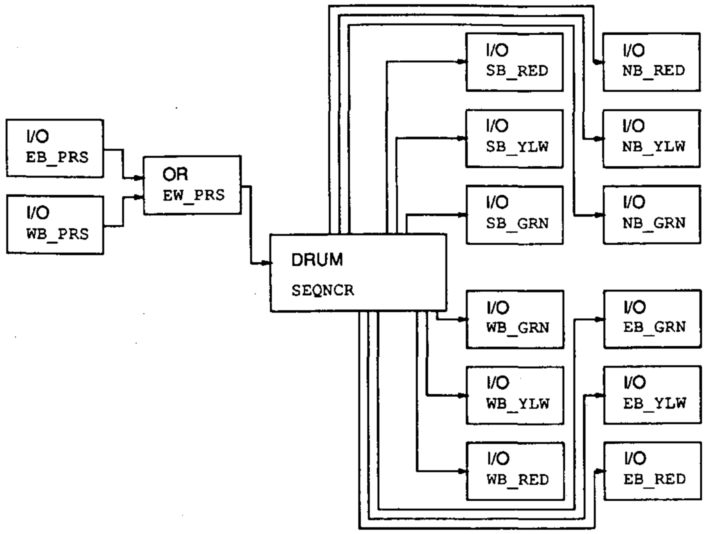


FIGURE 1 Semiactuated intersection control strategy using five function block types.

nections, or data flows, are shown as link connections between the boxes. A typical function block program resembles an activity-on-node (PERT) project management scheduling network.

Function block programming is different from the modular design taught in introductory programming classes because the end user never encounters any procedural code. All interaction with hardware devices, protocol conversions, buffers, timing demands, and error recovery are embedded in a parameterized function block graphical icon that can be configured by a traffic engineer using a function block editor. The blocks available within the function block editor are prepared by software engineers in a standardized manner, which permits seamless interconnection and implementation. This set of blocks is called TCBLKS, an acronym for traffic control blocks.

TRAFFIC ENGINEERING FUNCTION BLOCK PROGRAMMING MODEL

The previous section introduced the function block programming model. This section addresses three areas: (a) traffic engineering task vocabulary, (b) configuration of function block strategies, and (c) function block language structure.

Traffic Engineering Task Vocabulary

The set of building blocks available in the block library constitutes the "vocabulary" for users to assemble applications. Table 1 summarizes the 40 blocks that have been developed. This library includes signal sequencing blocks, signal filters, logic functions, interfaces to external sensors and actuators, archival functions, and various algorithmic blocks. The intent of establishing a definition of these control blocks is to provide a vocabulary that can be assembled by a traffic engineer (in a sketch or diagram) to define the required software. This concept is used extensively in the chemical and process engineering fields so that there is an almost one-to-one correspondence between the process and instrumentation diagram (P&ID) developed by the chemical engineer and a function block strategy constructed by the control system contractor. The same continuity is sought for traffic engineering.

Because this model is only in the prototype stage, the blocks described in Table 1 currently fall short of providing a comprehensive set of building blocks. To support the growth of this model, new blocks can be created and included in the block library as long as the new blocks conform to standard block definition and operation practices. Thus, applications such as a dynamic signal control algorithm such as OPAC (*6*) could be included in a single-function block. In general, this model supports blocks of varying execution complexity ranging from simple logic gates to complex blocks supervising several ramp meters. For example, the simple blocks, such as mathematical computations and digital logic, are necessary for incorporating minor operational changes typically required by peculiar geometric or policy constraints. In contrast, the complex blocks such as ramp metering or intersection control algorithms can provide rapid and reliable task-level programming.

Configuration of Function Block Strategies

An advantage afforded by the function block programming model and advocated in this paper is the ability to easily "program" or configure robust traffic control software without an extensive software engineering background. A typical configuration tool can operate like a simple vector drawing package commonly found on desktop or notebook computers. Instead of manipulating shapes and lines, it manipulates function blocks. A block program is developed by assembling a strategy composed of predefined blocks that provide common traffic engineering operations. The mechanics of constructing such a strategy can be viewed in three steps.

TABLE 1 Function Block Summary

AND, OR, XOR, NOT: These blocks perform the essential boolean logic operations on their input(s).

DlyOn, DlyOff, OneShot: These digital blocks perform digital logic timing operations. The DlyOn block delays a transition from low to high for a specified time. Alternatively, the DlyOff block delays a high to low transition for a specified time period. The OneShot provides a pulse generating mechanism for transitions from low to high.

D-Shift: provides a 16 bit shift register for transient storage of digital states.

D-UI: provides an operator with simple on/off and pulse operations for user interfaces.

Match: provides basic decoding functionality.

Timer: measures the duration of digital events.

Counter: can be used for counting lo-high transitions.

FF-RS, FF-D, FF-JK: These blocks provide discrete implementations of clocked RS, D, and JK flip flops. A T flip flop can be constructed from the JK flip flop.

Drum: provides state sequencing subject to minimum and maximum durations with the capability of back stepping.

Rate: calculates the filtered rate of an incoming digital pulse train.

Add, Mult, Div: These block provide basic mathematical operations.

Mavg, A-Shft: Both blocks implement a circular queue. The Mavg block uses the queue to compute the moving average of a time series. The A-Shft block provides a mechanism for introducing a time delay (lag).

A-Latch: latches an analog value when a digital pulse is received.

A-SWITCH: selects between two analog signals based on the state of a digital input.

A-UI: provides an operator with a mechanism to enter an analog value for a user interface.

Filter: provides a simple discrete approximation for a first order analog filter.

Test: compares an input against a set of absolute Hi and Lo bounds or relative to another signal. The results of these comparisons are digital points other blocks can connect to. It is useful for implementing conditional logic.

Sel-H, Sel-L, Sel-M: High, low and middle selector blocks. The first two blocks have two inputs, the middle selector requires 3 inputs.

RMSB: provides supervisory rate selection of a ramp metering rate based upon one upstream volume sensor and up to six downstream occupancy values.

LOOKUP: provides an interpolated lookup table for defining non-linear transformations

D-Coll, A-Coll: monitors up to eight inputs (Analog or Digital) and records their state to a file. A background spooler is set up so this file can reside on any OS-9 file device. These devices include hard disks, floppy disks, RAM disks and non volatile disks.

RMDI, RMDO: used to read digital inputs (DI) or write digital outputs (DO) on a 170 running ramp metering software.

RMRI, RMRO: used to read register inputs (RI) and write register outputs (RO) on a 170 running ramp metering software.

VMS: contains up to 8 prioritized ASCII messages that can be displayed on a variable message sign by a digital event.

1. Select: Blocks providing the requisite device interfaces, signal processing, control computations, cycle phasings, or data collection features are selected and placed on the drawing area.
2. Configure: Parameters defining a program block's operation, such as the number of phases or a loop detector's I/O port, are configured for each block. This procedure is performed by selecting a block with the mouse and choosing the "configure parameters" option. Of course, each block is instantiated with a full set of default values that may be acceptable, in which case this operation can be omitted for many blocks.
3. Connect: The blocks are connected by clicking on a block, selecting a particular block output connection, clicking on another block, and selecting a particular block input connection. Basic error checking is performed to prevent sockets with various data types from being connected. For example, it would be invalid to connect the state of a loop detector to the socket determining the cycle length for a traffic light drum sequencer.

These steps are intended only to give the reader an idea of how the function block model could be configured. In practice, these steps will likely be intertwined as a strategy is developed and edited incrementally. Past strategies would typically serve as templates for new applications. Also, a number of diagnostic, reporting, drawing, scaling, and annotating tools are necessary to round out the features of the configuration tool.

Language Structure

Table 1 provides a summary of some preprogrammed blocks that can used to develop block strategies. This section details the basic architecture of those blocks and how they can be assembled. Abstractly, a function block is a vector consisting of the following elements (Figure 2):

- Input sockets are used either to retrieve data from other blocks or are assigned constant values. Input sockets are actually references to memory locations from which the block reads values. The values stored in those locations can be changed either by another block's output socket or by an operator manually inserting a value. These sockets represent the destination half of a data flow connection.
- Local storage stores block parameters and interim calculations.
- Output sockets are used to store block output values and can be connected to other blocks. Output sockets are actually references to memory locations to which the block will write output data. The values written to those locations can be read by another block's input socket or by an operator examining sockets. These sockets represent the source half of a data flow connection.
- A block algorithm periodically reads the values associated with the input sockets, performs calculations, manipulates local storage, and then updates the output sockets.

Although blocks may have several input or output sockets, it is not required that they all be connected. In fact, input

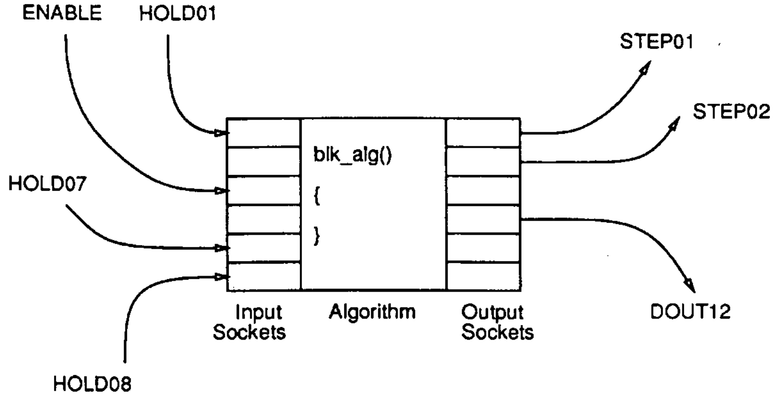


FIGURE 2 Function block components.

sockets can be assigned either constant values (Figure 3, Socket 3) or connected to another block's output socket (Figure 3, Socket 2) during configuration. Similarly, output sockets can be left dangling (Figure 3, Socket 1) or connected to input sockets (Figure 3, Socket 2) on other blocks. The only restriction on connecting blocks is that one input cannot be connected to more than one output socket (Figure 4).

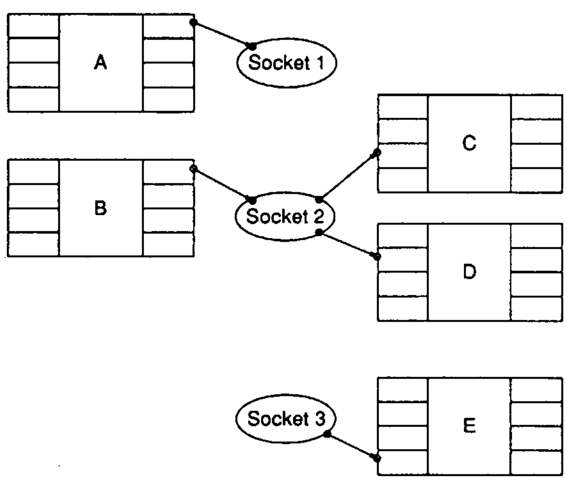


FIGURE 3 Example of valid socket connections.

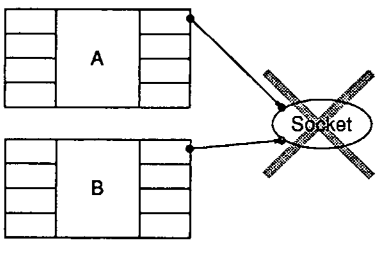


FIGURE 4 Example of invalid socket connection.

IMPLEMENTATION OF TCBLKS

Traffic engineers are likely to be most concerned with the block vocabulary, configuration concepts, and language structure of this traffic control software model. To round out the description of this function block model, a few important implementation concepts are addressed: (a) internal data model for the function blocks, (b) real-time scheduling, (c) capacity considerations, and (d) on-line user interfaces. Our purpose is not to formally define the model but to demonstrate an efficient real-time implementation and to give further insight into the software model. A more extended discussion appears in a previous paper (*7*).

Consider the example strategy shown in Figure 1. This strategy is composed of 16 blocks that describe which sensors should be read, which internal algorithms should be used, and which actuators should be manipulated. Without regard to how often the blocks must be run, this strategy can be described as a topologically sorted list of blocks to be run {EB__PRS, WB__PRS, EW__PRS, SEQNCR, SB__RED, . . . }. Each of these blocks must be represented internally as a data structure with local storage, input sockets, and output sockets. These data structures are different for each block type. For example, the DRUM block has 16 digital output sockets, but the OR block has only one digital output. To provide a structured method for interacting with the various data structures, a master list of blocks called the block table (Figure 5) maintains a list of all the symbolic block names and a code representing the class of blocks. For example, all OR blocks would have a class code of 11 and all I/O blocks would have a class code of 19. This code is used by the software model to determine which table to search to retrieve the data structure defining a block. For example, the table for OR blocks (Figure 5) would contain the data structures defining the EW__PRS block.

Connections between blocks are very important for this model because they provide the mechanism for communication. The connection table (Figure 5) provides a list of all data connections and includes the following information:

- Source socket is a symbolic name identifying the source of a data connection.

Block Table

Group	Block	Class

Connection Table

Source	Destin	Type	Index

OR, Class Code = 11

Group	Block	Local Storage	Input Sockets	Output Sockets

Digital Socket Table

Index	ddata	State

I/O, Class Code = 19

Group	Block	Local Storage	Input Sockets	Output Sockets

Analog Socket Table

Index	adata	State

DRUM, Class Code = 26

Group	Block	Local Storage	Input Sockets	Output Sockets

Text Socket Table

Index	tdata	State

FIGURE 5 Internal model for function blocks.

- Destination socket is a symbolic name identifying the destination of a data connection.
- Socket type indicates which table to look in for the socket. For the data model shown in Figure 5, this could be a reference to the digital, analog, or text socket tables.
- Socket index is used to locate the particular socket in a socket table (digital, analog, or text) identified by the socket type field.

In preceding sections, sockets have been conceptually diagrammed as tightly coupled with the block. However, to improve implementation efficiency, all sockets are stored outside the block and referenced via the connection and socket tables (Figure 5). There is a socket table for each possible data connection type. For example, digital states, analog values, or text messages are stored in the tables shown in Figure 5. When a block is executed, it references its input socket indexes (Figure 5) and retrieves the appropriate information from the socket table. Similarly, after the computations have been performed, it uses the output socket indexes to update the respective output sockets.

Real-Time Scheduling of Function Blocks

Because block processing is not instantaneous, the blocks must be scheduled such that all blocks have an opportunity to run often enough to meet their application requirements. One possible approach would be a round-robin scheduler. The problem with this type of scheduling is that when blocks are added or subtracted the timing characteristics change. This kind of side effect is unacceptable, particularly if interaction with a particular device or evaluation of a traffic signal phase change at regular intervals is necessary. A more sophisticated approach would be to run all the blocks at their fastest required rate (a least-common-denominator approach). This technique would be adequate if sufficient CPU cycles were available for executing all blocks at the fastest required rate. However, in practice, only a few blocks require very frequent service (say 50 Hz) and other blocks require service far less often (say 0.1 or 0.01 Hz).

Because of the varying timing requirements for different portions of a block strategy, it is desirable to be able to assign a processing period to a group of blocks. To provide this capability and introduce a hierarchical level of abstraction, blocks can be grouped and assigned a name and periodic execution rate (Figure 6). Two additional internal tables are constructed to maintain this information: the group table and the task table (Figure 7). An additional status field in the group table is used to turn on and off the processing for an entire group of blocks. From the user's perspective, a collection of groups assigned to periodic tasks constitutes an application program (Figure 8). In the application shown in Figure 8, the blocks in Groups A, B, and C would be run every T_1 sec. Similarly, the blocks in Groups D and E would be run every T_4 sec. Within each of these groups, the blocks, their type, and their configuration define the semantics of the application program.

To facilitate the orderly start-up and shutdown of a function block strategy, the software starts up in a single threaded mode. It reads the function block strategy, creates all the necessary data structures for execution, initializes all I/O devices, runs all blocks once to initialize them, spawns periodic tasks, and commences the periodic execution shown in Figure 8. The periodic tasks are created according to the period and priorities in the task table (Figure 7). Groups are assigned to these tasks according to the task field in the group table (Figure 7). When the software receives a signal to shut down, it allows the periodic tasks to complete their current cycle (only if block processing was in progress before the shutdown signal was received), returns to single-threaded operation, runs all blocks once (permits files to be closed and I/O to be left in a safe state), and then terminates. The state diagram for this behavior is shown in Figure 9.

Capacity Considerations

The periodic tasks shown in Figure 8 represent only one-half of the software model. In practice, interactions with I/O de-

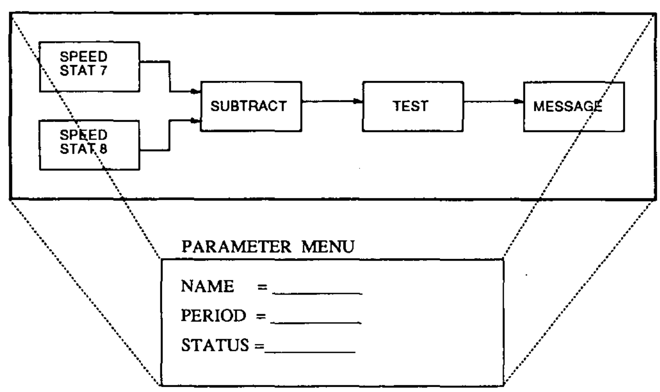


FIGURE 6 Block grouping illustration.

Task Table

Index	Period	Priority

Group Table

Group	Task	Status

FIGURE 7 Scheduling tables.

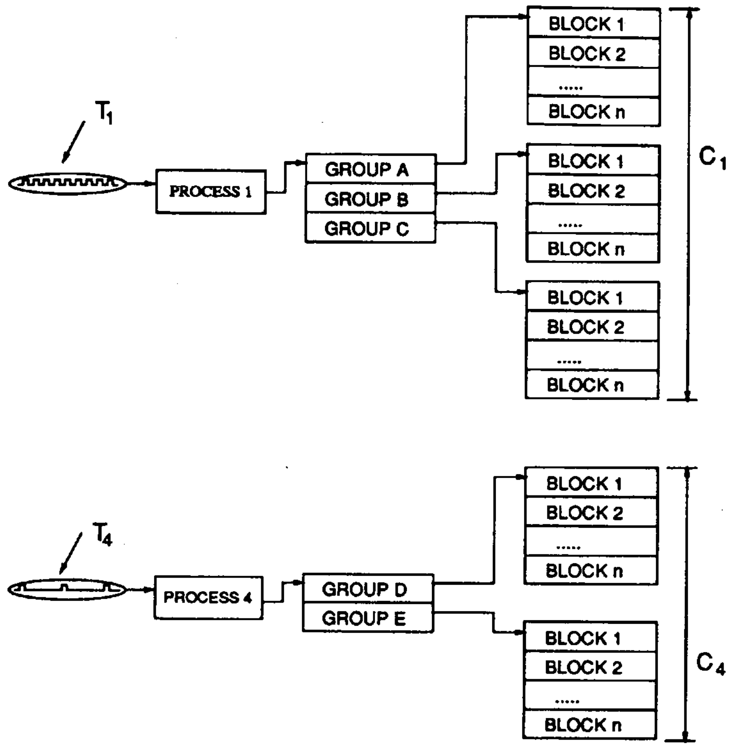


FIGURE 8 Processing of group and block structures by periodic task.

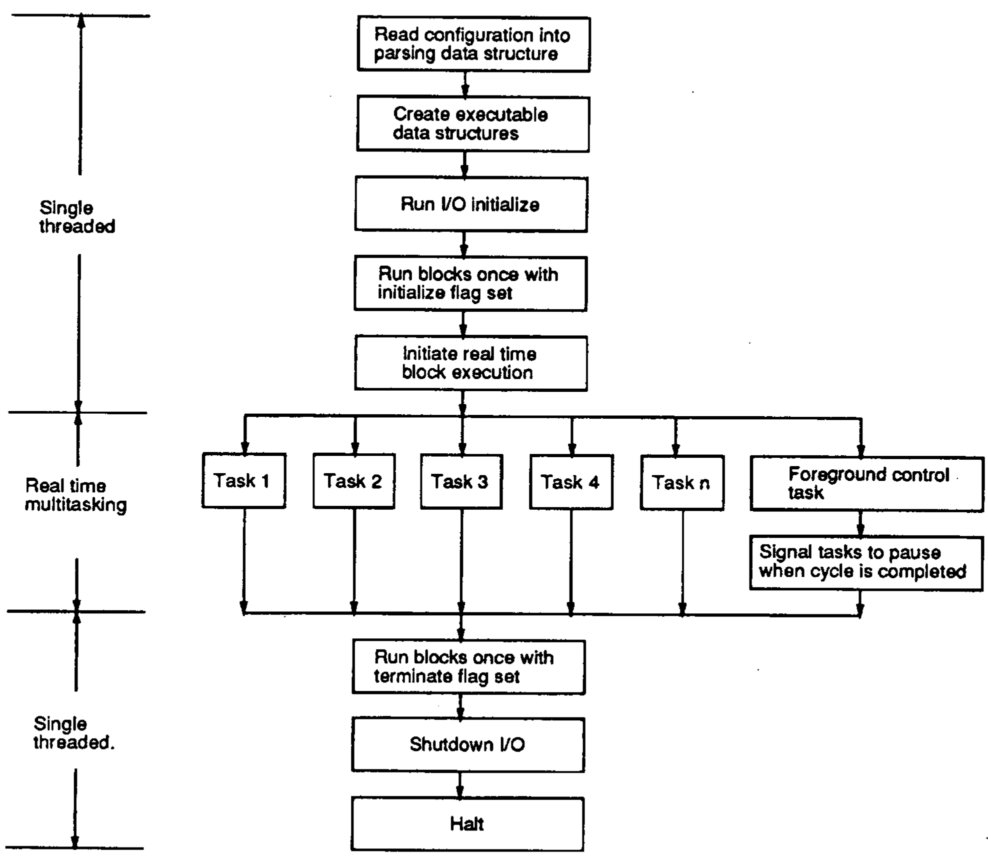


FIGURE 9 Software state diagram.

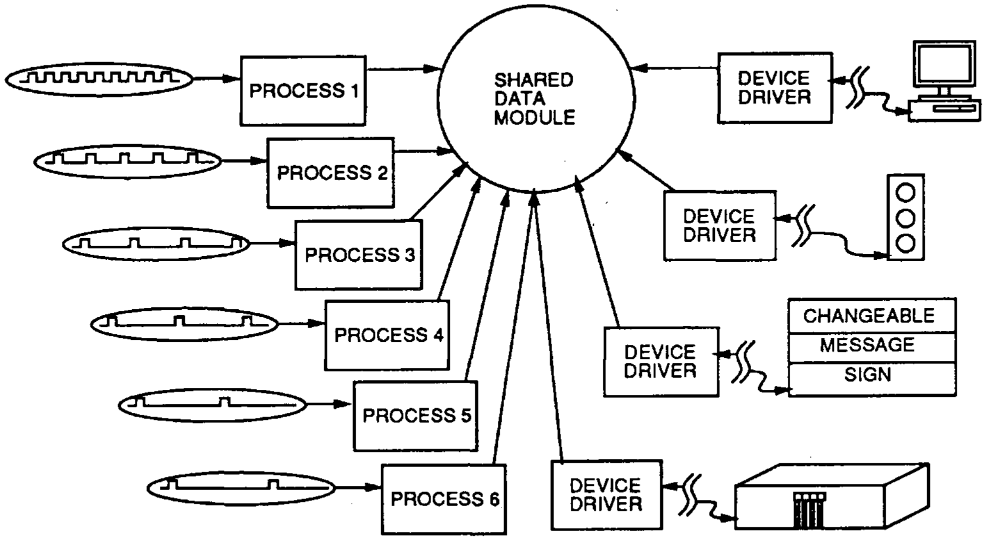


FIGURE 10 Task interaction.

vices such as serial ports, user interfaces, and disk drives have inherent time delays. To permit the processors to work on other duties, the periodic tasks do not directly interact with these devices. Instead, they communicate with asynchronous tasks using internal buffers. Conceptually, this software architecture looks like that shown in Figure 10. The periodic tasks that run the function blocks are shown on the left and the aperiodic tasks interacting with I/O devices are shown on the right. A complex set of tasks is shown in Figure 10. By inspection, it is not obvious whether the software model can respond to all the computational and I/O demands in a timely manner. For example, when monitoring loop detectors it is important that "passage pulses" not be missed. To guarantee that such events are not missed, it is necessary to determine whether the computational demand of the software (Figure 10) exceeds the capability of the processors. This evaluation can be performed using rate monotonic analysis techniques documented previously (*8–10*).

On-Line User Interfaces

The user interface for configuring the block strategy has been described in previous sections. The user interface for instrumentation and monitoring is also very important for development and diagnostic purposes. An interface such as the hex keypad and LED display found on the 170 or the alphanumeric display now being built into NEMA controllers could be used to interact with the ATC software. However, the function block model proposed in this paper provides a more intuitive method for interacting with the run-time control software. The basic concept for developing these "run-time user interfaces" is based on a client-server model in which the client is an operator interface program and the server is the function block processing program. Quite likely, the operator interface would be implemented on a notebook computer that could be plugged into a serial port on the ATC (Figure 11).

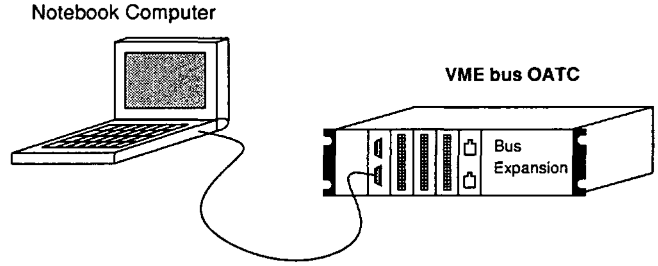


FIGURE 11 Use of a notebook computer to configure and monitor an ATC.

The client operator interface would interact with a strategy via the connection table and the various socket tables (Figure 5).

Interfacing with the controller in this fashion provides two important features. First, the client can symbolically reference any socket. So instead of the current practices on 170 controllers of looking at the word located at a particular hex offset, a symbolic name such as "Main&4th;NB_CNT.AOUT" could be used to read the volume counter on the northbound counter at Main and 4th. Second, the "State" field in the connection table restricts the ability of an operator interface program to write to a socket to only those input sockets not connected to other blocks (Figure 3, Socket 3). Of course, any point could be read by an operator interface, but unpredictable operation would result if an operator was trying to change an output socket that was also being changed by a function block (Figure 3, Sockets 1 or 2).

IMPLEMENTATION

The software model described in this paper has been implemented and tested in real time under simulated conditions for applications such as signalized intersections, ramp metering, and communication with existing traffic control devices. This software has also been used to implement a bottleneck surpervisory control strategy that was field tested along Highway 50 in Sacramento, California. The software communicated with Type 170 ramp meters over leased telephone lines and adjusted metering rates in response conditions at a downstream bottleneck. This demonstration was performed on the proposed Caltrans ATC platform configured with a 16-MHZ 68020 with 4 MB of RAM in November 1992.

ACKNOWLEDGMENTS

This work was supported in part by the California Department of Transportation under subcontract to the University of California Institute of Transportation Studies at Irvine.

REFERENCES

1. A. Rappaport and S. Haleri. The Computerless Computer Company. *Harvard Business Review*, Vol. 69, No. 4, July 1991, pp. 69–80.
2. M. J. Chase and R. J. Hensen. Traffic Control Systems—Past, Present and Future. In *Applications of Advanced Technologies in Transportation Engineering*, ASCE, Feb. 1989, pp. 257–262.
3. D. Bullock and C. Hendrickson. Advanced Software Design and Standards for Traffic Signal Control. *Journal of Transportation Engineering, ASCE*, Vol. 118, No. 3, May 1992, pp. 430–438.
4. M. Shaw. *Prospects for an Engineering Discipline of Software*. Technical Report CMU-CS-90-165. Carnegie Mellon University, Pittsburgh, Pa., Sept. 1990.
5. Quinlan, T. *Evaluation of Computer Hardware and High-Level Language Software for Field Traffic Control*. Technical Report. California Department of Transportation, Sacramento, Dec. 1989.
6. N. G. Gartner. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 75–81.
7. D. Bullock and C. Hendrickson. *A Model for Roadway Traffic Control Software*. Technical Report. Carnegie Mellon University, Pittsburgh, Pa., Dec. 1992.
8. J. Lehoczky, L. Sha, and Y. Ding. The Rate Monotonic Scheduling Algorithm: Exact Characterization and Average Case Behavior. In *Real Time Systems Symposium*, IEEE Computing Society, Dec. 1989, pp. 166–171.
9. J. Lehoczky, L. Sha, and J. Strosnider. Enhanced Aperiodic Responsiveness in Hard Real-Time Environments. In *Real Time Systems Symposium*, IEEE Computing Society, 1987, pp. 261–270.
10. C. L. Liu and J. W. Layland. Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment. *Journal of the Association for Computing Machinery*, Vol. 20, No. 1, Jan. 1973, pp. 46–61.

The views expressed by the authors do not necessarily reflect the individual views or policies of either the California Department of Transportation or the University of California.

Publication of this paper sponsored by Committee on Traffic Signal Systems.