Vehicular Grid Communications: The Role of the Internet Infrastructure*

Mario Gerla†, Biao Zhou†, Yeng-Zhong Lee†, Fabio Soldo*†, Uichin Lee†, Gustavo Marfia†
†Department of Computer Science, University of California, Los Angeles, CA 90095
*Dipartimento di Matematica, Politecnico di Torino, Torino, Italy 10129
{gerla,zhb,yenglee,uclee,gmarfia}@cs.ucla.edu, fabio.soldo@studenti.polito.it

ABSTRACT

Vehicle communications are becoming a reality, driven by navigation safety requirements and by the investments of car manufacturers and Public Transport Authorities. As a consequence many of the essential vehicle grid components (radios, Access Points, spectrum, standards, etc.) will soon be in place (and paid for) paving the way to unlimited opportunities for other car-to-car applications beyond safe navigation, for example, from news to entertainment, mobile network games and civic defense. In this study, we take a visionary look at these future applications, the emerging "Vehicular Grid" that will support them and the interplay between the grid and the communications infrastructure.

In essence, the Vehicular Grid is a large scale ad hoc network. However, an important feature of the Vehicular Grid, which sets it apart from most instantly-deployed ad hoc networks, is the ubiquitous presence of the infrastructure (and the opportunity to use it). While the Vehicular Grid must be entirely self-supporting for emergency operations (natural disaster, terrorist attack, etc), it should exploit the infrastructure (when present) during normal operations. In this paper we address the interaction between vehicles and Internet servers through Virtual Grid and Internet Infrastructure. This includes transparent geo-route provisioning across the Internet, mobile resource monitoring, and mobility management (using back up services in case of infrastructure failure). We then focus on routing and show the importance of Infrastructure cooperation and feedback for efficient, congestion free routing.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Network Architecture and Design, Routing protocols

*This work was supported in part by ONR MINUTEMAN grant N00014-01-C-0016.

WICON'06, The 2nd Annual International Wireless Internet Conference, August 2-5, 2006, Boston, MA, United States

General Terms

Design, Performance

Keywords

VANET, ad hoc networks, infrastructure support, geo routing, geo location service, load balancing

1. INTRODUCTION

Safe navigation support through wireless car to car and car to curb communications has become an important priority for car manufacturers as well as transportation authorities and communications standards organizations. New standards are emerging for car to car communications (DSRC and more recently IEEE 802.11p) [6]. There have been several well publicized testbeds aimed at demonstrating the feasibility and effectiveness of car to car communication safety; for instance, the ability to rapidly propagate accident reports to oncoming cars, the awareness of unsafe drivers in the proximity and the prevention of intersection crashes.

The availability of powerful car radios, and of abundant spectrum (when not used for emergencies) will pave the way to a host of new applications for the Vehicle Grid (V-Grid). These emerging applications span many fields, from office-on-wheels to entertainment, mobile Internet games, mobile shopping, crime investigation, civic defense, etc. Some of these applications are conventional "mobile internet access" applications, say, downloading files, reading e-mail while on the move, etc. Others involve the discovery of local services in the neighborhood (e.g., restaurants, movie theaters, etc.) using the vehicle grid as an ad hoc network. Others yet imply the close interaction among vehicles such as interactive car games.

To support the more advanced services, new brands of functions must be deployed such as creation/maintenance of distributed indices, "temporary" storage of sharable content, "epidemic" distribution of content and index. Examples include the collection of "sensor data" by cars as seen as "mobile sensor platforms," the sharing and streaming of files in a Bit-torrent fashion, and the creation/maintenance of massively distributed data bases with locally relevant commercial, entertainment and culture information (e.g., movies, hotels, museums, etc.). Typically, these applications are distributed and follow a P2P collaboration pattern.

One of the objectives of this paper is to investigate the impact of new P2P urban applications on V-Grid management and mobility management. More generally, we are interested in the functions and performance of the new network and transport protocols that support these P2P applications. Regarding support, one common assumption is that the basic V-Grid P2P protocols must be entirely self-supporting (no help from infrastructure). While this is important for emergency operations, where the infrastructure has failed (say, natural disaster, terrorist attack, etc.), in reality, the infrastructure will be accessible most of the time, and in fact it should be exploited in day-by-day operations.

This paper specifically addresses the interaction and co-existence of the P2P vehicle grid protocols with the Internet protocol stack, including the support of vehicle connections to and through the Internet. One important difference between Internet stack and V-Grid stack is the prevalent use of geo routing in the latter. We will investigate the issue of consistent addressing and transparent geo-routing when vehicle to vehicle connections are carried through the Internet. Related to addressing is the maintenance of a vehicle geo-location service both in the Internet and in the V-Grid, to be supported by a "vehicle overlay" within the wired Internet. Another important function provided by the infrastructure is the dynamic estimate of path quality to vehicles as well as the more general monitoring of congestion in the wireless V-Grid. These building blocks will collectively represent an essential support to efficient infrastructure and V-Grid designs.

This paper is organized as follows. In Section 2, we address the VANET environments and applications. In Section 3, we present the addressing in the Vehicular Grid. In Section 4, we detail our proposed geo location services. In Section 5, we propose a *hybrid* geo routing protocol that utilizes the Internet infrastructure as well as proposed geo location services. In Section 6, we evaluate our protocols through extensive simulations. In Section 7, we review the related work. Finally, we conclude the paper in Section 8.

2. VANET ENVIRONMENT AND APPLICATIONS

Vehicular networks provide a promising platform for future deployment of large-scale and highly mobile applications. With the increasing deployment of urban wireless access points, the application domain of traditional mobile and ad hoc networks (MANETs) is giving way to wireless mesh networks that extend wired line connectivity to the Internet. Vehicular ad hoc networks (VANETs), however, remain a largely unexplored platform for a class of compelling applications. Given the automobile's role as a critical component in peoples' lives, embedding software-based intelligence into them has the potential to drastically improve the user's quality of life. This, along with significant market demand for more reliability, safety and entertainment value in automobiles, has resulted in significant commercial development and support into deployment of vehicular networks and applications. In this section, we outline key differences that distinguish the vehicular platform, introduce applications by their interactions with data, and describe a number of constraints and challenges for the vehicular application infrastructure.

2.1 VANET environment

In designing protocols for the next generation vehicular network, we recognize that nodes in these networks have significantly different characteristics and demands from those in traditional wireless ad hoc networks deployed in infrastructureless environments (e.g. sensor field, battlefield, etc.). These differences have significant impact on application infrastructures. First, automobiles have much higher power reserves than a typical mobile computer. Power can be drawn from on-board batteries, and recharged as needed from a gasoline or alternative fuel engine. Second, automobiles are orders of magnitude larger in size and weight compared to traditional wireless clients, and can therefore support significantly heavier computing (and sensorial) components. This combined with plentiful power means vehicular computers can be larger, more powerful, and equipped with extremely large storage (up to Terabytes of data), as well as powerful wireless transceivers capable of delivering wire-line data rates. Third, automobiles travel at speeds up to one hundred miles per hour, making sustained, consistent vehicle-to-vehicle communication difficult to maintain. However, "existing statistics" of vehicular motion, such as tendencies to travel together or traffic patterns during commute hours, can help maintain connectivity across mobile vehicular groups. Finally, vehicles in a grid are always a few hops away from the Infrastructure (WiFi, cellular, satellite, etc.). Thus, network protocol and application design must account for easy access to the Internet during "normal" operations. In this paper, we also foresee the value of the vehicle grid as an emergency network when the infrastructure fails. We must therefore design protocols and applications that survive (with possible degraded performance) when isolated from the Internet.

2.2 Emerging applications

Another important departure of vehicle networks from conventional ad hoc networks is the opportunity to deploy, in addition to traditional applications, a broad range of innovative content sharing applications (typically referred to as Peer-to-Peer applications). While their popularity has been well documented, they have been thus far confined to the wired Internet (e.g., BitTorrent). The storage and processing capacity of VANET nodes make such applications feasible. Moreover, the fact that car passengers are a captive audience provides incentive for content distribution and sharing applications at a scale that would be unsuitable to other ad hoc network contexts. Therefore, it is important to understand the role of the V-Grid in these applications. Here, we describe a representative set of VANET P2P applications and classify them by the vehicle's role in managing data: as a data source, data consumer, source and consumer, and intermediary.

First, the vehicle provides an ideal platform for mobile data gathering especially in the context of monitoring urban environments (i.e., vehicular sensor networks) [16, 17]. Each vehicle can sense events (e.g., images from streets or the presence of toxic chemicals), process sensed data (e.g., recognizing license plates), and route messages to other vehicles (e.g., forwarding notifications to other drivers or police officers). Because vehicular sensors have few constraints on processing power and storage capabilities, they can generate and handle data at a rate impossible for traditional sensor networks. These applications requires persistent and reliable storage of data for later retrieval. In addition, they require networking protocols (including sophisticated query processing) to efficiently locate/retrieve data of interests (e.g., finding all the vehicles at a certain time and location).

Second, the vehicles can be significant consumers of contents. Their local resources are capable of supporting high fidelity data retrieval and playback. For the duration of each trip, drivers and passengers make up a captive audience for large quantities of data. Examples include locality-aware information (map based directions) and content for entertainment (streaming movies, music and ads) [24, 25]. These applications require high throughput network connectivity and fast access to desired data.

In a third class of compelling applications, vehicles are both the producers and consumers of content. Examples include services that report on road conditions and accidents, traffic congestion monitoring, and emergency neighbor alerts, e.g. my brakes are malfunctioning [5, 23]. We note that their direct relevance to road safety makes them a high priority for commercial entities. These applications require real-time, location-aware data gathering/dissemination and retrieval.

Finally, all of the above applications will need to rely on vehicles in an intermediary role. Individual vehicles in a mobile group setting can cooperate to improve the quality of the applicant experience for the entire network. Specifically, vehicles will provide temporary storage (caching) for others, as well as forwarding of both data and queries for data. In this capacity, they require reliable storage as well as efficient location of and routing to data sources and consumers.

The demands of these applications give us a list of requirements and challenges for vehicular applications. Note that we can leverage them to simplify the applications infrastructure.

- *Time sensitivity* – Time-sensitive data must be retrieved or disseminated to the desired location within a given time window. Failure to do so renders the data useless. This mirrors the needs of multimedia streaming across traditional networks, and we can leverage relevant research results from the related areas.
- *Location awareness* – Both data gathered from vehicles and data consumed by vehicles are highly location-dependent. This property has direct implications on the design of data management and security components. Data caching and indexing should focus on location as a first order property; while data dissemination must be location-aware in order to maintain privacy and prevent tampering.

3. ADDRESSING IN THE VEHICULAR GRID

A major challenge in the management of vehicular network mobility and the interconnection of vehicles to and through the Internet is addressing. Let us begin with two important definitions:

- *Unique car ID* – This ID is immutable for a practically long time. Several options are possible here: license plate number, vehicle ID, owner's name, etc. IP address is not immutable, but can occasionally be used as "temporary" unique ID.
- *Routable car address* – The routable address is the address that contains all the information required to deliver the packet to destination, with proper support in the network. For example: geo-coordinates; a specific attribute (as in "attribute based" routing - e.g., TorrentID in CarTorrent [24]); unique ID (IP or MAC address) for some type of routing, e.g., AODV or DSR.

We envision that the addressing scheme should support the following services:

- It must enable each car to efficiently address any other car in the V-Grid.
- It must enable an Internet server to address any vehicles in the grid.

One can easily see that the geo address (coupled with georouting) fits these requirements. More precisely, geo-routing dispatches a packet in proximity of the target destination. Once the packet is within radio reach, any unique node identifier (e.g., license plate number) can be used to deliver the packet to the node. The geo address continuously changes as the node moves. Thus, it is the task of the Geo Location Service to provide the up-to-date index to the location. Beside geo routing, AODV can also be very effective over short paths (few hops), using IP addresses as routable addresses to set up/maintain on demand routes. However, AODV rapidly degrades with number of nodes and speed; it is not suitable for V-Grid operations.

Because Internet type IP routing (i.e., prefix routing, etc.) is based on the Internet hierarchy, it is extremely inefficient in a highly mobile ad hoc network where the hierarchy must be reconfigured every time node connectivity changes. Yet, the IP address is still useful as an identifier in AODV routing and in TCP connections. For these reasons, we propose to maintain in the vehicle grid a "unique" IP address for cars.

The technique currently used in wireless LANs and Mesh Networks to dynamically assign a different IP address within each Access Point domain guarantees uniqueness only in a limited scope. It is not adequate for global V-Grid uniqueness. We assume there will be no DHCP for cars when they pass by Access Points. We propose that the car IP address be initialized, for example, by hashing license plate number, vehicle ID. With very small probability, the resulting IP address may not be unique. If address conflicts ever happen (during TCP connection set up, or AODV routing), the tie is broken by re-hashing the IP address (i.e., the IP address of a car may change during lifetime). To avoid collisions in a proactive way, IP uniqueness within a local scope (say, 3-4 hops) may be constantly verified and enforced by an elected IP master node.

4. GEO LOCATION SERVICE

As mentioned earlier, a critical component of the georouting address structure is the Geo Location Service (GLS) – a distributed service that maps any car ID to the set of most recent geo locations. We propose to implement two "parallel" versions of GLS, namely: OLS (Overlay Location Service) and VLS (Vehicle Grid Location Service). OLS is maintained within the Internet infrastructure; VLS is maintained entirely in the vehicle grid. The two services are synchronized, but are independently maintained to provide fault tolerance (one of the tenets of our proposed architecture). The rationale for the redundant implementation is to assure a backup in the V-Grid when the infrastructure fails.

Figure 1 illustrates a possible OLS implementation. An overlay structure is established in the urban Internet. Each car, as it passes by Access Points, periodically (say every minute) registers its ID (uniquely identified by hashing, for example, license plate number, vehicle ID, etc.) and the current geo-location. OLS maintains an index of IDs. Each

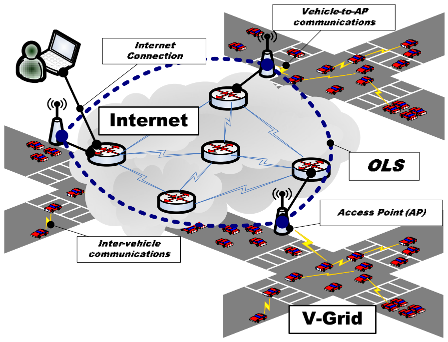


Figure 1: Vehicle grid extension example

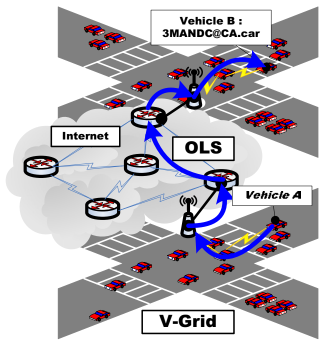


Figure 2: Routing in the Overlay

ID is mapped to the most recent geo coordinates (thus allowing motion prediction). The index is distributed across the overlay. It may be managed via DHT (Distributed Hash Table).

To illustrate the OLS operation, suppose that host A (fixed or mobile) wants to establish a TCP connection to mobile host B as shown Figure 2. Host A first queries OLS with *3MANDC@CA.car* starting from the nearest server of the OLS "overlay." It gets back the "recent" geo-locations, the IP address, and other additional information of host B, through which it can predict the more recent location of the car. As a result, it derives the best AP to reach the destination. Host A sends the message to that AP; the message is encapsulated, e.g., in an IPv6 network envelope that contains the geo address in the extended header. Routing in the Overlay is based on geo address. Namely, the destination geo address determines the AP at the end of the Internet path, before the V-Grid segment. At destination, the AP geo-routes the packet to the ad hoc network with both geo address and car ID in the header. The car responds with own IP address and geo address and directs the response (encapsulated in the overlay envelope) to the sender IP address. Note that the encapsulation into a geo routed network envelope is identical regardless whether the sending Host A is fixed or mobile.

In principle, one geo location server, say OLS, would be sufficient. However, since the urban Internet infrastructure (or wireless access to it) may fail, we also maintain VLS in the Urban Grid. In normal operations, VLS can be simply synchronized with OLS. However, since VLS must perform also in isolation, a distributed implementation is in order.

Considerable amount of research has been done on distributed Geo Location Servers. One must minimize the update overhead incurred as the car moves and at the same time minimize the time to search the distributed index – two conflicting requirements. Geo Location Services, based on multi-tier geographical hashing, are proposed in [19, 27], which are known to perform better than flat schemes.

In this paper we propose to implement VLS as a modification of the HIGH GRADE scheme [27]. Like HIGH GRADE, VLS is based on a hierarchy. The main difference with respect to HIGH GRADE is that at the lowest level, there is a unique (mobile) geo server – e.g., a CalTran truck – that roams in a cell ($1km \times 1km$), periodically advertising its coordinates. To make the service more robust, multiple servers may coexist in a cell, one being elected by lowest ID. In HIGH GRADE, in contrast, all cars are potential servers. Here, the cars in the "perimeter" enclosing the hashed coordinates (i.e., location server points) all must store the mapping. This causes problems when the cars move. They must hand off the maps that have become "off-center." In our proposed VLS scheme, cars register locally with the truck as part of the lowest service tier. The VLS truck never hands off as long as it remains in the cell. The periodic advertisements from the truck make it easy to route the updates to the truck without incurring the well known dead-end problems of geo routing. Advertisement rate and associated overhead can be kept low exploiting Directional Forward routing [18]. At the higher tiers of the hierarchy VLS is identical to HIGH GRADE. There may be as many as 1000 cells in a large metropolitan area (say $33km \times 33km$). A car will geo-hash its license number, for example, in one of these cells. As the car moves to the next cell, it registers with the local truck. At the same time, it updates the server in the next level up to point to the new cell. It is important to note that during normal operations, when the infrastructure is up and accessible, the OLS servers are also updated in synchrony with VLS updating. In fact geo server updating can be also done through the Internet overlay, eliminating multi-hop geo routing through the V-Grid and reducing the VLS overhead.

5. INTERNET INFRASTRUCTURE SUPPORT

From the description of the V-Grid architecture in the preceding sections, we can now summarize the main benefits deriving from the coexistence of V-Grid and Infrastructure:

1. Support transparent geo location service and geo routing. This is achieved with an overlay that hosts the OLS (Overlay Location Server). The OLS is synchronized with VLS and helps reduce VLS update overhead.

2. Reroute car to car connections via the wired internet, when it is more cost-effective than routing them directly on the V-Grid

3. Monitor V-Grid data traffic conditions; detect and report incipient congestion to the APs and indirectly to the V-Grid. Congestion reports by AP's to V-Grid nodes permit the implementation and execution of efficient Call Admission Control policies

4. Assist in reporting "V-Grid" time variable connection capacity to Internet servers. This information complements capacity measurements collected directly by Servers end-to-end (e.g., via CapProbe [9]). It allows for smooth vertical handoff across different wireless media. It also paves the way for rate adaptation,

format/content adjustment and possible transcoding through the Overlay proxies.

The Infrastructure also stands to benefit from the ubiquitous presence of the V-Grid, specifically in the cases of Infrastructure overloads, failures and total collapse due to natural disasters (e.g., earthquakes, floods) or terrorist attacks. In the latter case, smooth switchover from operational mode to "attack" mode can position the V-Grid as a viable, robust (albeit capacity limited) backup to the urban Infrastructure.

5.1 Rerouting and load balancing

In the remaining of this section we address the problem of optimal re-routing of car to car traffic through Access Points and wired backbone. The problem is to select the best of two paths: the path entirely contained in the ad hoc network, and the path that uses the wired Internet. It is clear that packets between opposite sides of the city will be routed via the Infrastructure rather than on wireless paths involving many hops. Likewise, "gossip" packets seeking cooperating neighbors one or two hops away will be routed directly. There is a "grey zone" between these extremes where one must carefully trade off number of hops and load on the two paths. Associated with routing problem is "call admission control." When both the ad hoc path and the wired Internet path are congested, new connections must be rejected and on going connections must be dropped starting from the lower priority ones.

For robustness we are seeking a distributed solution to the routing and load balancing problem. Most of the distributed solutions proposed so far optimize geometric properties such as minimum path length, or min hop pair of disjoint paths, or max bandwidth path. Very few schemes are adaptive to load. AODV has been extending to finding the path with max load threshold, or delay threshold. However, the AODV procedure requires flooding. Flooding is exactly what georouting is trying to avoid. Proactive schemes such as OLSR can propagate periodic updates with load information associated to "nodes and links." But, this periodic updating would introduce undesirable overhead in the vehicular network. For the vehicular network we propose a hybrid approach that exploits the wired Infrastructure. Namely, each source makes a distributed decision using "hints" provided by the infrastructure Overlay. The scheme works as follows:

1. The Vehicle Overlay (VOL) collects and stores the information about loads measured at all the AP's in the V-Grid.

2. The vector of AP loads is broadcast by each AP into its area of influence (using scoped flooding).

3. The source estimates the hop distance to destination and to AP's based on source/destination coordinates, city map and radio transmission range.

4. Using the "load" hints received from VOL, the source can then compute a combined hop and load cost for both direct path and AP routed path.

In essence, the wireless source estimates shortest paths using geo coordinates, but needs VOL assistance to incorporate loads in the cost criteria. Note that, using VOL, the cost to measure and distribute load information to wireless nodes is minimal. The same information would be prohibitively expensive if propagated using exclusively the ad hoc network. Note also that the same load information can be used for admission control, by discarding packets and rejecting connections based on priorities and congestion thresholds.

5.2 Optimal multicommodity flow routing

In this section we attack the rerouting and load balance problem as a multicommodity flow (MCF) problem under a carefully chosen utility function. The function must capture the tradeoff between the benefit of getting this flow through versus the cost of increasing the congestion in the network. This type of problem has been efficiently solved in wired packet networks using a centralized approach [1]. The wireless broadcast channel renders the formulation much more complex than in the wired network case. Moreover, it would be a formidable challenge to apply the centralized approach to a dynamic ad hoc network like the vehicular network, because of number of problems including: collecting the source/destination traffic requirements; delivering this data to a central server; making the optimal routes available with low latency to sources before transmission, and; assuring that the data is not obsolete in the face of rapidly moving nodes and changing requirements. In this section, we present a non-linear programming model to achieve the optimal routing that minimizes delays in wireless ad hoc network with infrastructure. From this, we show upper bounds on performance and verify the tradeoffs and trends derived from simulations.

It is important to note that we must account for the interference among wireless nodes to use MCF in wireless environments. Jain et al. [8] used the concept of a clique in the contention graph to represent interference for a given topology graph. A similar approach was used in [3]. This approach is impractical for a large network since it implies the knowledge of all cliques in the contention graph that takes an exponential time. [10, 11] proposed a simple geometric method to infer interferences called Transmitter-Receiver Conflict Avoidance (TRCA); i.e., if node i is within node j's transmission range and node i is transmitting, then node j must be silent during the time for node i to finish transmitting. Our model uses this approach. Figure 3 illustrates the contention region for the link (i, j). Since many source-destination pairs can potentially use link (i, j) we impose a "bundle" constraint for each link $(i, j) \in E$ such that the whole flow through the contention region of the link (i, j) must not exceed its capacity C_{ij}. It is important to note that all the previous works have failed to clearly identify how interferences affect the capacity of a wireless link and what are the similarities and differences between wired and wireless multicommodity flow. To the best of our knowledge, this is the first attempt to correct this problem.

Let us consider a network with a set of nodes $N = N_A \cup N_I$, where N_A represent the set of ad hoc network nodes, and N_I represent the set of APs. Let $E = E_A \cup E_I$ be set of links where $e_a \in E_A$ represent a wireless link and $e_i \in E_I$ represent a wired link. Each link $(i, k) \in E_A$ has an associated a capacity C_{ik}. For simplicity we assume that the capacity on wired link is infinity.

Let x^k_{ij} denote the amount of flow of commodity k on link (i, j) and let $\underline{x}^k$ and $\underline{b}^k$ denote, respectively, the flow vector

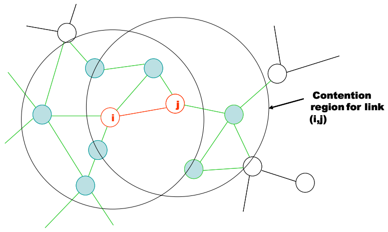


Figure 3: Contention region for the link $(i,\, j)$

and demand vector for commodity k. The ordinary mass balance constraints can be written as

$$\mathrm{N}\underline{x}^k = \underline{b}^k \tag{1}$$

where N is the incidence matrix for graph $G(N, E)$. As shown in Figure 3, we consider the total flow through the contention region, since this represent the total amount of traffic that contend for link $(i,\, j)$. Let $S_{ij} = \sum_{k \in K}(x^k_{ij} + x^k_{ji})$ be the total amount of flow on link $(i,\, j) \in E_A$. This can be formulated as follow:

$$\sum_{k \in K} x^k_{ij} + I_{ij} \le C_{ij} \tag{2}$$

Here, we define interference on link $(i,\, j)$ as follows.

$$\begin{aligned} I_{ij} = \sum_{k \in K} x^k_{ji} + \sum_{l \in N} \sum_{m \in N} \Gamma_{ijl}(1 - \Gamma_{ijm}) S_{lm} \\ + \sum_{l \in N} \sum_{\substack{m \in N, m > l \\ (l,m) \neq (i,j)}} \Gamma_{ijl}\Gamma_{ijm} S_{lm} \end{aligned} \tag{3}$$

where $\Gamma_{ijn} = 1$ iff n interfere with i or j; otherwise $\Gamma_{ijn} = 0$. Equation 2 can be simplified as

$$\sum_{k \in K} x^k_{ij} \le \tilde{C}_{ij} \tag{4}$$

where $\tilde{C}_{ij} = C_{ij} - I_{ij}$.

This shows that wireless MCF flow can be formulated in the same fashion as wired MCF. Note that unlike wired networks where the capacity is always fixed, in the wireless network, the interference reduces the actual capacity. We can interpret this result by introducing a *reduced capacity*, $\tilde{C}_{ij}$, function of routing variables and C_{ij}. If there is no flow (trivial case) or if the active links are "far enough" one from the other so that they do not interfere when they transmit, then $C_{ij} = \tilde{C}_{ij}$. In all others cases we have $C_{ij} > \tilde{C}_{ij}$.

The formal analogy with the well known MCF[1] also stimulate research to investigate how classical algorithms for the wired MCF can adapt to the wireless case; for example, how to solve this model with a decentralized approach, or how to extend it to the multicast case. Note also that Equation 4 is a necessary condition for a flow to be feasible. In addition, these constraints are tighter than others existing constraints. For instance, it is easy to prove that if Equation 4 holds, then node interference constraints introduced in [13] also hold.

To compare analytic results with simulation results we used a load balancing criterion that minimizes the maximum delay (based on M/M/1 approximations):

$$\min \left\{ \max_{(i,j) \in E_A} \frac{f_{ij} + I_{ij}}{C_{ij} - (f_{ij} + I_{ij})} \right\} \tag{5}$$

[1]Note that in classical MCF the bundle constraints write as $\sum_{k \in K} x^k_{ij} \le C_{ij}$

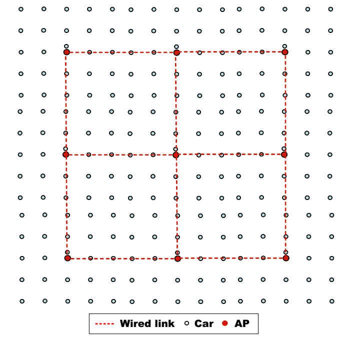


Figure 4: Topology

where $f_{ij} = \sum_{k \in K} x^k_{ij}$, the total flow from on link $(i,\, j)$. Then we evaluate the average end-to-end delay obtained with this optimization criterion.

Equation 5 subject to Equation 1, Equation 4 and $\underline{x}^k \ge 0$ is a convex optimization problem. We use `cvx` [4] to solve the problem. In our implementation, we also include the packet flow due to the 802.11 acks to make it more closer to the simulation scenarios. We use a path-flow formulation with variables $f(P^k_h)$, $k = 1, \cdots, K$, and $h = 1, 2, 3, \cdots, M_k$. Thus, we have

$$x^k_{ij} = \sum_{h=1}^{M_k} f(P^k_h)\delta^k_{ij} \tag{6}$$

where P^k_i is the i-th path from s_k to d_k and $\delta^h_{ij} = 1$ iff $(i,\, j) \in P^k_h$, 0 otherwise. Since simulation deals with geo-routing, we allow only two paths to be feasible for each commodity, i.e., the shortest path in ad hoc network and the shortest path using infrastructure. For implementation details, readers can refer the extended version of this paper [7].

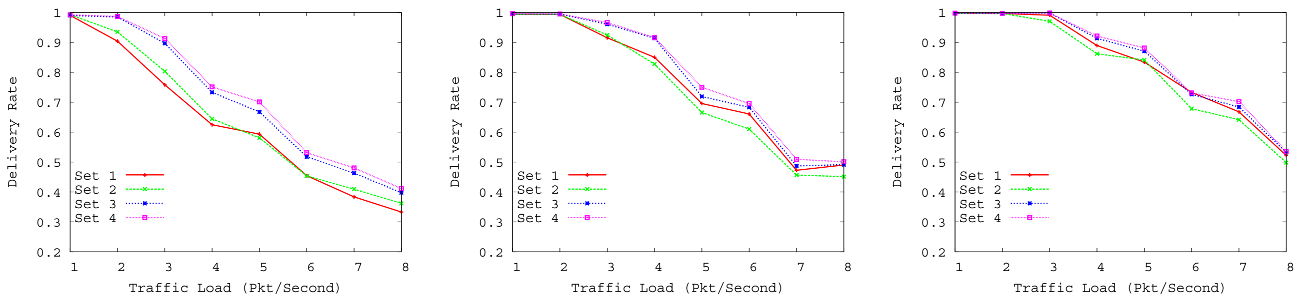
6. EXPERIMENTAL EVALUATION

In this section we evaluate the efficacy of the infrastructure support in managing vehicular communications. In particular we are focusing on the routing of traffic between vehicles. The basic goal is to select the best of two options: routing on the grid and; routing via the nearest Access Points through the wired Internet Infrastructure. These two options and their implementations have been discussed in Section 5. A key feature of our proposed Vehicular Overlay architecture is the load information offered to vehicles by proxy servers. In the experiments we will assess the improvement achieved when this feature is offered.

6.1 Experiment setting

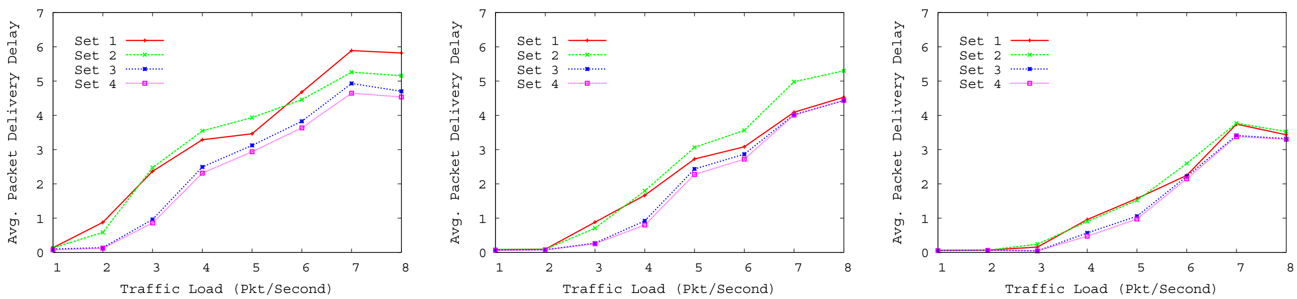
The experimental setting consists of a $1.5km \times 1.5km$ Manhattan type grid with 225 nodes regularly placed at street intersections, as shown in Figure 4. In the grid there are horizontal, vertical and also diagonal streets. Nine Access Points provide access to the Internet Infrastructure and to the Global Internet. The radio transmission range is 175m, thus each vehicle has eight neighbors. Also, any vehicle is at most 2 hops away from the nearest AP. Channel data rate is 2Mbps. IEEE802.11b MAC protocol is assumed.

In our experiments we first consider a static scenario in order to exclude the perturbations introduced by mobility, for example, "voids" and disconnections in the topology. These motion induced effects would inject "noise" in the load balancing assessment. The traffic pattern consists of 200 con-



(a) $f_{ap} = 0.50$ (b) $f_{ap} = 0.75$ (c) $f_{ap} = 0.90$

Figure 5: Delivery Ratio vs. Traffic Load.



(a) $f_{ap} = 0.50$ (b) $f_{ap} = 0.75$ (c) $f_{ap} = 0.90$

Figure 6: Average Data Packet Delay vs. Traffic Load (Static)

nections. A fraction f_{ap} of the connections is between vehicles and Internet servers (e.g., web accesses, file downloads, videoconferencing, etc.) (i.e., through Access Points). The remaining $1 - f_{ap}$ fraction is car to car (e.g., navigation safety, traffic information, forensic messaging, etc.). We consider in our study both UDP and TCP connections, in separate experiments. UDP connections carry CBR traffic (e.g., multimedia). TCP connections carry bulk data traffic. The connections inject traffic with variable offered rate (from 1 to 8 $pkts/sec$). Both TCP and UDP connections have 66.6% duty cycle, i.e., they are up 66.6% of the time, and they are dormant the remaining 66.6% of the time. Thus, 133 connections are up on average. Total simulated time is 300s.

To evaluate the importance of balancing loads in the vehicular network, we carry out the following sets of experiments:

- Set 1: All the traffic goes through the nearest access point and via the Infrastructure, including car to car traffic.
- Set 2: The fraction f_{ap} of the traffic goes through the nearest access point; the remaining $1-f_{ap}$ fraction (the car to car traffic) goes on the Grid path, regardless the length of the connection (in hops) and the traffic type.
- Set 3: The fraction f_{ap} goes to nearest AP; the remaining fraction (car to cat traffic) uses the min hop option between the Grid path and the AP+Internet path.
- Set 4: Same as Set 3, except that the selection between Grid path and AP+Internet path is determined by a combination of hop distance and AP link congestion.

The UDP results are reported in Figure 5 and Figure 6. Both delivery rate and delay are reported with $f_{ap} = 0.5$, 0.75 and 0.9. We note that the performance ranking of the solutions is quite consistent throughout the experiments. As expected the load balanced solution (Set 4) gives the best delivery ratio; the solution that routes all traffic to the AP's (Set 1) does worst in most cases. In fact, AP's are the bottlenecks in this system. The delays are consistent with delivery ratios; the higher the drop, the higher the delay. We also note that as the fraction f_{ap} increases, i.e., more and more traffic are directed to the Internet, the delivery ratio improves (to 100% for $f_{ap} > 0.75$) and the delay decreases from several second to milliseconds. This is easily explained by noting that when all the traffic goes to the internet, the traffic pattern exhibits localities at the APs resulting in excellent spatial reuse. When the traffic is mostly car to car with a fair number of long range connections, the spatial reuse is less prevalent and the network becomes quickly saturated, with high drop rates, heavy queues and delays up to several seconds.

This trend is clearly visible from the results. As the fraction f_{ap} increases from 0.50 to 0.90, i.e., as more as more traffic is to/from the Internet, the performance improves. At 0.75 and 0.90 all the schemes guarantee full delivery for offered rate $= 2$ $pkts/sec$. Also, the gap between the solutions decreases. This does not mean that the shortest path and load balancing solutions are not effective in that region. Rather, it means that the contribution to the total improvement is less significant. Recall that at $f_{ap} = 0.90$, car to car traffic is only 10% of total traffic.

The TCP experiments for $f_{ap} = 0.5$ are reported in Figure 7 and are compared with the UDP results. At $f_{ap} = 0.5$, the network becomes congested already at 2 pkt/sec, thus load backs up at the TCP connections creating an "infinite backlog" condition. Consequently, TCP throughput performance is constant for varying offered load. The first observation is that Set 4 (load balancing) performs best - with highest throughput. This was as expected. However, the throughput improvement is minimal. Several factors may

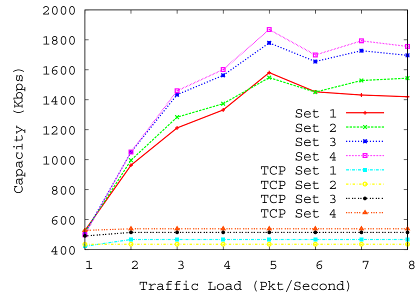
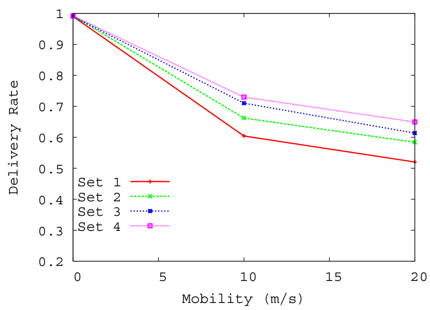
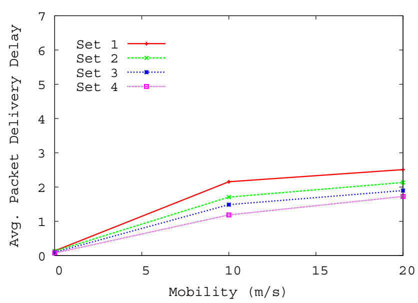


Figure 7: Capacity vs. Traffic Load for both UDP and TCP (f_{ap}=0.5, Static)



(a) Delivery Ratio vs. Mobility.



(b) Average Data Packet Delay vs. Mobility.

Figure 8: Impact of mobility (f_{ap}=0.5)

play here. One is the fact that at $f_{ap} = 0.5$ the AP's are heavily loaded and load balancing will seek grid paths with potentially more hops. However, more hops mean higher packet drop rate for TCP, which implies window smaller than optimum and suboptimal throughput. As a result, the benefit of a less congested path is counteracted by the problem of multiple hops, with little benefit deriving from load balancing. Another interesting observation is that TCP throughput is only about 1/4 of UDP throughput at saturation. In other words, TCP can exercise congestion control and keep the network lightly loaded, but at a high performance cost. It might be advisable to seek other methods for efficient congestion control in the vehicle Grid.

The previous experiments were based on a static topology. The reason for a static topology was to separate the load and path length contributions from the motion effects. Now, since we have shown the superiority of the path and load aware routing scheme, we wish to validate this result also when nodes move. Indeed, Figure 8 confirms the efficacy of rerouting for speeds up to 20m/s (over 70 mph). Nodes move according to the Random Way Point model. The Geo Routing scheme suffers a severe throughput degradation due to mobility, about 40% at 10 mph. This behavior is well known and has been documented by several other researchers. Interestingly, by comparing Figure 8(a) with Figure 5(a) we note that this is the same degradation suffered, under similar traffic pattern, in the static network when the load is increased from 1 $pkts/sec$ (stable condition) to 4 $pkts/sec$ (severe overload).

Equally surprising is the fact that the load balancing approach yields a 20% throughput improvement here as it did for the overloaded network. This is in part explained by the fact that mobility reduces geo-routing efficiency as suitable forwarding neighbors are often missed. This reduction in capacity can be in part recovered by load balancing on less loaded paths. The delay behavior, shown in Figure 8(b), is consistent with throughput results. In summary, rerouting is cost-effective also when nodes move, and the ranking of rerouting schemes is identical as the one established for static nodes.

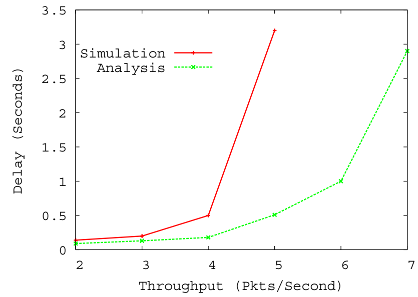


Figure 9: Delay vs. throughput – simulation vs. multicommodity optimization

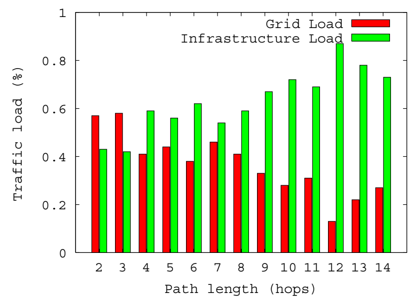


Figure 10: Infrastructure vs Grid routing for different Source-Destination hop distance

6.2 Multicommodity flow model results

The multicommodity flow optimization was applied to the same configuration shown in Figure 4 in order to derive optimistic upper bounds on throughput performance. The results are shown in Figure 9 and Figure 10.

Figure 9 shows that the optimal solution (with very optimistic assumptions) leads to a limiting throughput that is about 0.5 higher than simulation results. Clearly, further work on the model is required to reduce the gap between model and simulation. Once we are confident in the "realism" of the multicommodity flow results, we can shift attention to the actual load balance algorithm seeking improvements to close the gap from that side. Figure 10 shows that as the hop distance between Grid source-destination pair, a progressively larger fraction of the load is shifted to the wired infrastructure. This confirms the findings of the simulation experiments.

6.3 Lessons learned from the experiments

The set of rerouting and load balancing experiments was rather preliminary; yet, one can already draw some lessons and directions for future study and experimentation. In particular:

- If the vehicular traffic is mostly to AP's and Internet servers, the system behaves like a "cellular" system, with good spatial reuse. The use of sophisticated routing and load balancing is not critical.
- As the traffic shifts from a mostly AP pattern to a Car 2 Car pattern the importance of load balancing and shortest path routing becomes more pronounced. The shift from AP bound to C2C traffic may be caused, as we conjectured, by the increasing popularity of P2P applications – CarTorrent, Internet games, safety navigation, mobile sensor platforms, etc. – and by the shift in the location of resources – from Internet Servers to vehicles.
- The necessity for efficient distributed congestion control in the V-Grid is quite apparent, especially when distributed car to car communications will become prevalent. TCP does not seem to be very effective in this camp, as it reduces traffic to 1/4 of achievable capacity. More work is needed on TCP over the Grid. Another direction to explore is Connection Acceptance control. Low priority applications may be controlled or even disabled in heavy traffic situations. The Vehicle Overlay may effectively assist vehicles in performing global, network wide congestion control.
- Effective admission control and congestion control will be mandatory in order to support interactive applications such as safe navigation, VoIP, videoconferencing, remote collaboration. Delays will need to be kept in the 100 ms range or less. Again, the Vehicular Overlay will be effective in this direction.
- Mobility causes throughput degradation across all rerouting schemes, but, at the same time confirms the importance of load balancing across the infrastructure
- The multicommodity flow models show a rather large gap between simulation and analysis. This discrepancy will be further investigated, leading to more realistic multicommodity models, but also to better performing load balance algorithms.

7. RELATED WORK

The benefits of using the infrastructure to enhance the per node throughput capacity of an ad hoc network are well documented in [14, 20]. The asymptotic capacity is found to be $\Theta(\sqrt{N/\log N})$-fold better than in a flat ad hoc network, mainly due to the fact that relaying over APs effectively reduces the mean number of hops from source to destination.

Most proposed hybrid (infrastructure + ad hoc) networks are to provide extended coverage of existing services, e.g., wireless LAN and 3G, without however considering the rerouting of ad hoc packets over the infrastructure. Lee et al. [15] proposed multi-hop extension of the wireless LAN. Their work is focused on increasing the coverage and performance as well as allowing incremental deployment; thus, a multi-hop path is dynamically managed using passively measured bandwidth and latency. Luo et al. proposed an architecture called UCAN [21] which utilizes ad hoc routing over 802.11-based interfaces to improve the performance of 3G cellular networks. Similarly, a Wireless Mesh Network (WMN), which is composed of a set of fixed nodes with multiple wireless interfaces, is being deployed as a solution to extending the reach of the last-mile access to the Internet, using a multi-hop ad hoc routing [12].

In addition to the extended coverage, a hybrid network can also route packets to a remote mobile node over the infrastructure, which is the main focus of this paper. Miller et al. [22] proposed a hybrid network implementation where an AP restricts its wireless multi-hop service to k-hops for efficiency. A mobile node joins an AP after receiving a beacon by setting up a default path to that AP. The overall routing protocol shares the same idea of AODV, but their protocol allows the route discovery packet to travel over the APs. Moreover, mobile IP for ad hoc networks is used to support communication with hosts on the Internet. WIANI [2] extended the previous approach and pursued load balancing among APs. To this end, an AP periodically sends a beacon with "local" load information; from this information, a mobile node makes a probabilistic join decision only if it finds a AP with load below a given threshold.

A common problem with the previous approaches is that they propose to use RREQ flooding over APs to select and set up the rerouting path. Unfortunately, this does not scale as the number of mobile nodes and APs increases. In this paper, we differentiate ourselves from previous approaches in that we propose an Overlay Location Service (OLS) solution to the problem. OLS maintains geographic locations of APs and mobile nodes, and allows mobile nodes to efficiently utilize geo-routing not only over the vehicular grid but also over the Internet. In addition, OLS provides a "global" view of AP congestion levels, thus leveraging efficient use of communication resources.

A scheme closer to OLS is the use of a "mobile" UAV based airborne backbone to interconnect teams operating in a tactical environment [26]. The issue of rerouting from ad hoc ground paths to airborne paths via UAVs is similar to the issue of rerouting from Grid to Infrastructure. However, the ad hoc nature of the airborne "infrastructure" and the presence of teams in the battlefield (as opposed to individual cars in the vehicle grid) leads to an integrated routing solution (Landmark routing) not feasible in the urban vehicle environment. Moreover, the airborne UAV backbone poses a number of additional problems such as dynamic optimization of UAV number and position, and possible "delay tolerant" interconnects of isolated ground teams which do not apply to the urban wired infrastructure.

8. CONCLUSIONS

In this paper we have proposed a novel Vehicle Grid architecture that includes the wired backbone Infrastructure as a key element to provide Grid operation support (traffic management, mobility, etc.), and; that provides a routing strategy for the wireless ad hoc network. A prominent feature of the Infrastructure is the Vehicle Overlay. We have shown that the overlay can efficiently support routing propagation across the urban environment and can assist the V-Grid in traffic rerouting when the ad hoc path is not cost-effective, in congestion control and in vertical handoff.

A number of significant research challenges remain to be explored, including:

- Time-sensitive dissemination of data to and from vehicles.
- Efficient data indexing and querying mechanisms using location and secondary characteristics.
- Reliable and persistent network-based storage in the presence of node churn (movement in and out of local mobile groups) and unreliable hardware.
- Reliable location-based communication in the presence of high vehicular mobility, intermittent connectivity and lossy channels.

These problems are the focus of ongoing research in our group.

9. REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, 1993.

[2] J. Chen, S. Li, S.-H. G. Chan, and J. He. WIANI: Wireless Infrastructure and Ad-HocNetwork Integration. In *IEEE ICC'05*, Korea, May 2005.

[3] L. Chen and S. H. Low. Joint Congestion Control and Media Access Control Design for Ad Hoc Wireless Networks. In *IEEE INFOCOM'05*, Miami, FL, USA, Mar. 2005.

[4] CVX: Matlab Software for Disciplined Convex Programming. `http://www.stanford.edu/~boyd/cvx/`.

[5] M. D. Dikaiakos, S. Iqbal, T. Nadeem, and L. Iftode. Vitp: An information transfer protocol for vehicular computing. In *ACM VANET'05*, Cologne, Germany, Sept. 2005.

[6] Standard Specification for Telecommunications and Information Exchange Between Roadside and Vehicle Systems - 5 GHz Band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Sept. 2003.

[7] M. Gerla, U. Lee, B. Zhou, Y. Lee, G. Marfia, and F. Soldo. Vehicular Grid Communications: The Role of the Internet Infrastructure. Technical report, UCLA CSD, 2006.

[8] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu. Impact of Interference on Multi-hop Wireless Network Performance. In *ACM MobiCom'03*, San Diego, CA, USA, Sept. 2003.

[9] R. Kapoor, L.-J. Chen, L. Lao, M. Gerla, and M. Y. Sanadidi. Capprobe: A simple and accurate capacity estimation technique. In *ACM SIGCOMM'04*, Portland, OR, USA, Aug.-Sept. 2004.

[10] M. Kodialam and T. Nandagopal. Characterizing Achievable Rates in Multi-hop Wireless Networks. In *ACM MobiCom'03*, San Diego, CA, USA, Sept. 2003.

[11] M. Kodialam and T. Nandagopal. The Effect of Interference on Capacity of Multi-hop Wireless Networks. Technical report, Bell Labs, Lucent Technologies, 2003.

[12] M. Kodialam and T. Nandagopal. Characterizing the Capacity Region in Multi-radio Multi-channel Wireless Mesh Networks. In *ACM MobiCom'05*, New York, USA, Jan. 2005.

[13] V. Kolar and N. B. Abu-Ghazaleh. A Multi-Commodity Flow Approach for Globally Aware Routing in Multi-hop Wireless Networks. In *PerCom'06*, Pisa, Italy, Mar. 2006.

[14] U. C. Kozat and L. Tassiulas. Throughput Capacity of Random Ad Hoc Networks with Infrastructure Support. In *ACM MobiCom'03*, San Diego, Jan. 2003.

[15] S. Lee, S. Banerjee, and B. Bhattacharje. The Case for a Multi-hop Wireless Local Area Network. In *IEEE INFOCOM'04*, Hong Kong, Mar. 2004.

[16] U. Lee, E. Magistretti, B. Zhou, M. Gerla, P. Bellavista, and A. Corradi. Efficient Data Harvesting in Mobile Sensor Platforms. In *IEEE PerSeNS Workshop*, Pisa, Italy, Mar. 2006.

[17] U. Lee, E. Magistretti, B. Zhou, M. Gerla, P. Bellavista, and A. Corradi. MobEyes: Smart Mobs for Urban Monitoring with a Vehicular Sensor Network. Technical Report 060015, UCLA CSD, 2006.

[18] Y.-Z. Lee, M. Gerla, J. Chen, J. Chen, B. Zhou, and A. Caruso. Direction Forward Routing for Highly Mobile Ad Hoc Networks. *Ad Hoc and Sensor Wireless Networks Journal*, 19(12), Jan. 2006.

[19] B. Li, J. Jonnotti, D. D. Couto, D. R. Karger, and R. Morris. A Scalable Location Service for Gographic Ad Hoc Routing. In *ACM MobiCom'00*, Boston, MA, USA, Aug. 2000.

[20] B. Liu, L. Z, and D. Towsley. On the Capacity of Hybrid Wireless Networks. In *IEEE INFOCOMM'03*, San Francisco, CA, USA, Mar. 2003.

[21] H. Luo, R. Ramjee, P. Sinha, L. Li, and S. Lu. UCAN: A Unified Cellular and Ad-Hoc Network Architecture. In *ACM MobiCom'03*, San Diego, CA, USA, Jan. 2003.

[22] M. J. Miller, W. D. List, and N. H. Vaidya. A Hybrid Network Implementation to Extend Infrastructure Reach. Technical report, UIUC CSD, Jan. 2003.

[23] T. Nadeem, S. Dashtinezhad, C. Liao, and L. Iftode. TrafficView: Traffic Data Dissemination using Car-to-Car Communication. *ACM Mobile Computing and Communications Review (MC2R)*, 8(3):6–19, July 2003.

[24] A. Nandan, S. Das, G. Pau, M. Sanadidi, and M. Gerla. Cooperative downloading in Vehicular Ad Hoc Networks. In *WONS'05*, St Moritz, Switzerland, Jan. 2005.

[25] A. Nandan, S. Das, S. Tewari, M. Gerla, and L. Klienrock. AdTorrent: Delivering Location Cognizant Advertisements to Car Networks. In *WONS'06*, Les Menuires, France, Jan. 2006.

[26] K. Xu, X. Hong, and M. Gerla. Landmark Routing in Ad Hoc Networks with Mobile Backbones. *Journal of Parallel and Distributed Computing (JPDC)*, 63(2):110–122, 2003.

[27] Y. Yu, G.-H. Lu, and Z.-L. Zhang. Enhancing Location Service Scalability with HIGH-GRADE. In *IEEE MASS'04*, Fort Lauderdale, Florida, USA, October 2004.