OCEAN – Scalable and Adaptive Infrastructure for On-board Information Access

Boualem Benatallah* Mahbub Hassan* Lavy Libman† Aixin Sun*

*School of Computer Science and Engineering
University of New South Wales
Sydney, NSW 2052, Australia
{boualem,mahbub,aixinsun}@cse.unsw.edu.au

†National ICT Australia
Bay 15, Australian Technology Park
Eveleigh, NSW 1430, Australia
Lavy.Libman@nicta.com.au

UNSW-CSE-TR-0601
January 2006

THE UNIVERSITY OF
NEW SOUTH WALES

SYDNEY · AUSTRALIA

Abstract

The idea of providing seamless connectivity and information access to users on-board public transport vehicles has attracted increasing popularity in recent years, as is evidenced by several commercially available systems that have attempted to implement it. In this article, we overview the specific technological challenges, research issues, as well as opportunities, that arise in the context of providing communication and information access on public transport. We focus on both the *networking* perspective — in particular, discussing extensions required to existing TCP/IP mechanisms to support the moving on-board networks — and the *data management* perspective, e.g. personalization of information and caching/pre-fetching for a highly dynamic and heterogeneous population of users. We contend that, to offer the best performance and flexibility, the design of public transport information systems ought to take advantage of the synergy between these two areas.

Index Terms

On-board networks, public transport, network mobility.

I. INTRODUCTION

It is an undeniable fact that mobile wireless communications have enjoyed a phenomenal success, with both voice and, more recently, data communication services becoming indispensable for an ever-growing number of people. This applies in particular to passengers on public transport vehicles, such as regular commuters in metropolitan areas or long-distance travelers, for whom the ability to communicate and access information while on the move is crucial to maintain productivity or provide entertainment during travel. However, with very few exceptions, currently available mobile communication systems do not offer any special support to public transport passengers; rather, the mobile end users simply connect via a wireless link directly to the static infrastructure (e.g. a base station in a cellular system). As a result, the communication abilities of mobile users are, by and large, limited to accessing services and applications that are common in the static world, such as phone calls and Internet access.

However, the vision of a truly ubiquitous mobile connectivity goes far beyond a mere last-hop access extension, and includes a wide variety of scenarios based on communications other than simple client access of a static network. The case of public transport in particular gives rise to a broad range of new applications that are tailored to its nature. These include, for instance, on-board entertainment that involves communications with a content server or among passengers, broadcast of on-board information, and more specialized applications such as mobile telemedicine (i.e. remote treatment of emergency on-board medical situations). The only communication option that is currently available in most instances, where each individual passenger is left to use her own connection (usually to a cellular infrastructure), cannot adequately and efficiently support this variety of applications. Moreover, public transport communications need not necessarily involve the passengers only; other, and no less important, applications may require communicating with mounted devices in the vehicle or with its driver/operator. Examples include real-time on-board surveillance, remote telemetry of vehicles in a fleet (e.g. monitoring of fuel gauges), and communications among drivers and/or between drivers and devices mounted on the roadside (such as cameras) about traffic congestion and hazards. Clearly, to support such a wide range of application requirements efficiently, it is imperative that alternative, non-traditional communication solutions become widely deployed on public transport vehicles.

Specifically, the communications architecture we discuss in detail in this article is depicted in Figure 1. The end users simply plug into a high-speed *on-board local-area network* (OBLAN), which communicates with the rest of the world as a single aggregated mobile entity, facilitated by a *mobile router* (MR). Thus, the wireless link is no longer limited merely to providing last-hop access for individual users; rather, it provides a flexible environment in which on-board users (including passengers and fixed vehicle-mounted devices) have the choice of communicating either among themselves, with fixed on-board servers, with peers in neighboring vehicles, or to the Internet.

The increasing popularity of the OBLAN architecture is evidenced by a growing number of commercial systems and research projects [1]–[6]. Furthermore, the Network Mobility (NEMO) working group of the Internet Engineering Task Force (IETF) have proposed an extension of the mobile IPv6 protocol to support mobile networks [7].[†] However, notwithstanding the initial availability of commercial solutions, much research is still required in order to realize the full potential of the architecture and deploy scalable and flexible information access systems in public transport. This article describes critical technological challenges and research issues that lie ahead in the quest to enable the variety of new applications and optimize the network performance and quality of service perceived by the on-board users. The article's content is a result of our activities in Project OCEAN (On-board Communication, Entertainment, And informatioN) [8], conducted at the University of New South Wales, Sydney, Australia. We provide an overview of our work and refer to other papers for detailed descriptions of the different concepts and techniques. Due to lack of space, we omit a detailed discussion of related work.

In the rest of this article, we first elaborate on the technical advantages of the on-board LAN architecture, as opposed to the current paradigm of individual communications. We then describe the characteristic features of public transport and OBLANs, and discuss how these features give rise to new challenges, yet also present new opportunities that have not been possible for individual users. Broadly, the research issues in the context of on-board networks can be categorized into *networking* and *data management* problems. In networking, much needs to be done towards extending TCP/IP mechanisms to support on-board users seamlessly and efficiently. Data management issues include caching, pre-fetching, and personalization of information in a highly dynamic user environment. As

[†]We note that additional communication solutions have been proposed to further support specific sectors of public transport (e.g., deployment of repeaters along railway tracks). Such solutions do not conflict with, and in fact are complementary to, the OBLAN architecture.

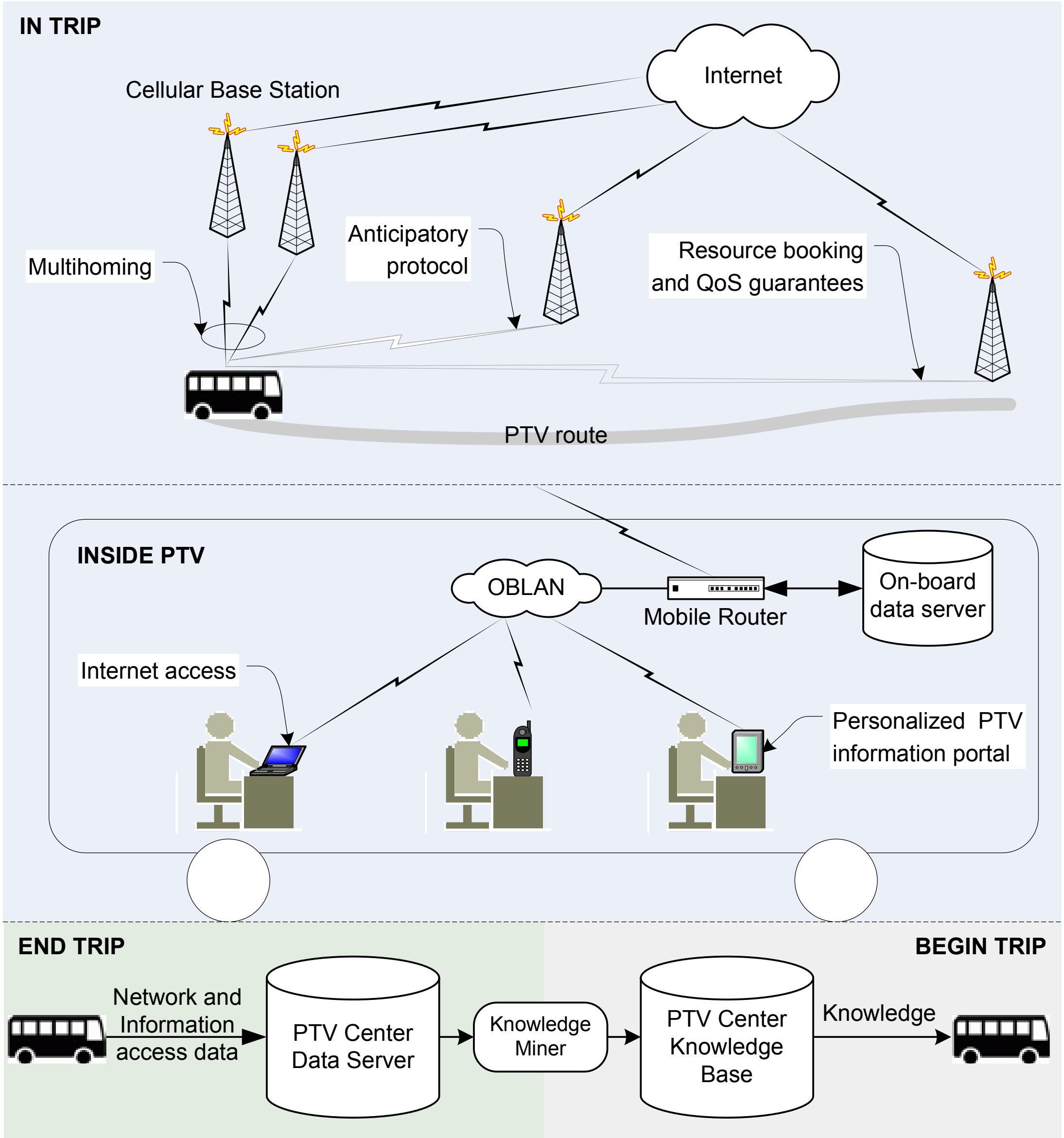

Fig. 1. On-board communication architecture.

we shall see, the synergy between these two disciplines is crucial to exploit the full benefits of OBLANs, in particular by capitalizing on the repetitive and predictable nature of public transport routes and passengers.

II. Advantages of On-Board LANs

As outlined in the introduction, the vast majority of mobile communication systems today still follow the *individual approach*, where each passenger is required to maintain its own wireless connection to a provider network (Figure 2a). For example, a bus passenger may carry a laptop equipped with a 3G wireless card, and connect to a nearby 3G base station without requiring any on-board infrastructure support. Due to its current availability, it is expected that this communication paradigm may remain the dominant method of on-board information access in the next few years. Before discussing the advantages of alternative architectures, it is important to understand the inherent limitations of the individual approach, which make it unsustainable for on-board communications in the long run:

- **Limited scope**. User devices are usually equipped with wireless cards with limited communication range, which renders them useless in airplanes, ships, or in public transport vehicles (PTVs) travelling outside the

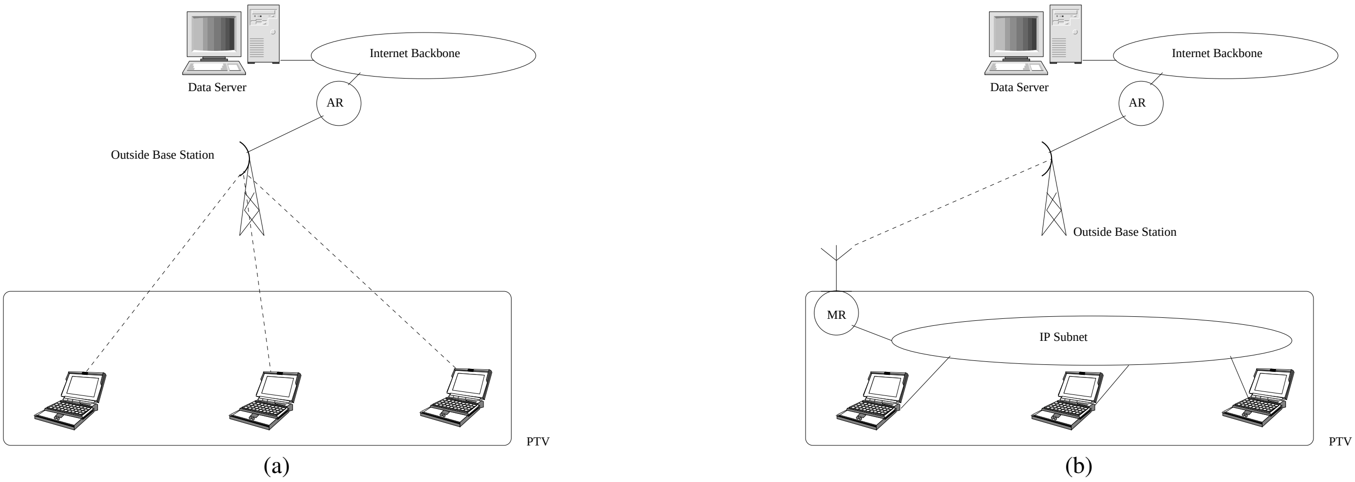

Fig. 2. Comparison between (a) individual approach and (b) on-board LAN approach.

coverage area of ground communication infrastructure (e.g. intercity trains). In such cases, each user would need to supply the equipment for direct communication with a satellite, making it impractical for typical users.

- **Limited bandwidth**. In the typical case (for terrestrial PTVs), where the individual connections are made to cellular networks, the bandwidth accessible to the individual user may not be adequate for many applications. For example, the 3G standard promises only 144 Kbps connections to moving users. Some high-end multimedia applications cannot be used over such a low access bandwidth.
- **Local on-board communications.** The need to involve the wide-area wireless infrastructure for all communications, even to fellow passengers on the same vehicle, renders the individual approach unsuitable for applications requiring only local on-board interaction, whether between peer passengers (e.g. multi-player games) or with an on-board server (e.g. for entertainment or travel information). Apart from the problem of limited bandwidth, the perceived latencies will be typical of a wide-area scale, and the cost may be prohibitive as well.
- **Scalability**. In order to maintain continuous connectivity, passenger devices must be directly connected to a nearby base station. When a PTV moves out of one base station's coverage area and enters another, all active individual sessions must perform a handoff (disconnect from the old base station and reconnect to the new one). Furthermore, all devices, whether active or not, must update their location so that future calls can be delivered to them. The number of simultaneous handoffs a base station has to accomplish is linear in the number of active users in the PTV. Also, the frequency of such handoffs increases with the speed of the PTV. Clearly, this creates serious scalability problems as more on-board passengers choose to access on-line information, manifested by the need to reserve significant extra capacity or suffer from an increased rate of disconnections.

In the on-board LAN approach (Figure 2b), additional networking infrastructure is deployed on-board the PTV, consisting of an on-board LAN (OBLAN) and a mobile router (MR) that allows the OBLAN to access a mobile wireless Internet connection (MWIC) [9]. The OBLAN provides a local high-speed connectivity to the on-board passengers (and other devices integrated in the vehicle). Hence, the wireless link is no longer the last hop in the end-to-end connections.[†] The MWIC may be implemented, e.g., over a land-based cellular station in urban areas (as depicted in Figure 2b), a satellite (in the countryside, at sea, or in the air), or potentially over a multihop ad-hoc network that involves other neighboring MRs along the route.

This infrastructure eliminates the need for individual users to maintain direct links with wide-area wireless providers; rather, all passenger devices merely connect to the OBLAN, and maintaining connectivity to the wide-area network is the responsibility of the MR. This addresses the limitations of scope and cost (e.g., the equipment for satellite communication need not be present on each end device), as well as scalability (e.g., only one mobile

[†]This does not mean that the OBLAN itself cannot be wireless, e.g. deployed using IEEE 802.11 technology; indeed, this is quite likely especially for public transport routes where the user population is very dynamic, such as short-haul bus trips. However, the *mobility* of the LAN is shielded from the users, who connect to it just like any stationary wireless LAN.

entity must be tracked by the ground cellular network). The bandwidth problem is addressed by making use of high-speed LAN technologies on-board the PTV, as well as provisioning the MWIC with a larger bandwidth than is normally allocated to an individual user. Since the bandwidth pool is shared by all users (and not hard-divided among them as in the individual approach), it becomes possible to take advantage of burstiness in the user data traffic and support a larger number of simultaneous connections (the *statistical multiplexing* effect).[‡]

Beside overcoming most of the individual approach limitations, the mobile LAN architecture brings about several additional benefits:

- **Communication reliability**. Inside a PTV, the wide-area wireless signal strength is diminished considerably by the metal shield of the vehicle, while competing mobile devices of other users generate interference. On the other hand, in an on-board LAN architecture, the MWIC connecting the on-board network can make use of an antenna mounted on the vehicle exterior (or even several antennas, taking advantage of multiple-input multiple-output (MIMO) technology); furthermore, the transmission power is virtually unlimited, which is not the case for the small and battery-powered user devices. As a result, the superior signal-to-interference/noise ratio (SINR) leads to reduced bit error and/or packet loss rates and lower communication delays.
- **Power consumption**. The fact that user devices need to communicate only locally, rather than with a remote wide-area network infrastructure, translates to a significantly reduced power consumption and consequently to a prolongued battery life.
- **Reduced bandwidth waste**. It can be expected that passengers travelling together in a PTV are likely to access the same popular content items, such as news, travel information, *etc*. With an on-board LAN architecture, such items can be cached in an on-board data server and then reproduced locally, instead of having each user retrieve them over and over again. Likewise, real-time streaming content (e.g. an Internet radio station) need only be transferred via the MWIC once and can then be multicast on-board.
- **Ease of upgrading**. Whenever a new wide-area communication technology (e.g. WiMax) or an improvement to existing protocols comes along, upgrading the equipment in a relatively small number of PTVs can be achieved far more easily and quickly than waiting for the upgrades to be gradually adopted by the general public. The same can be said about various other kinds of updates (e.g. security patches against cellular viruses).

III. System Characterisitics

The discussion above has demonstrated the great promise of mobile network systems based on the OBLAN architecture. However, before the benefits of such systems can be fully realised, it is imperative to appreciate the unique features exhibited by OBLAN-based systems and resolve the challenging problems, as well as capitalize on the new opportunities, that these features present. In this section we describe several characteristics that set apart OBLAN-based communications from traditional, personal mobile communications based on the individual approach, and discuss critical research issues that arise as a result, setting the stage for the subsequent description of the research directions tackled in the OCEAN project [8].

A. Aggregated mobility

The most obvious feature of mobile communications in PTVs is the fact that mobility is *aggregated* — in other words, a large number of users travel together, with an identical location and trajectory, and appear to the rest of the Internet as a single mobile subnetwork. This feature raises a number of complications. First, an OBLAN that aggregates tens (or potentially even hundreds) of users requires a sizeable amount of bandwidth and other resources from the wide-area network. Most existing wide-area wireless technologies, e.g. cellular, are designed under a paradigm that assumes a population of small users with random and independent mobility patterns. Since this assumption is not applicable to OBLANs, network operators may fear an increased likelihood of disruptions, e.g. when OBLANs are handed over between cells, and therefore be reluctant to allocate the same total amount of

[‡]We note that other, non-TCP/IP on-board infrastructures have been suggested in the past; see, e.g., [10], [11] for a *mobile cell* approach, where a base station is deployed on-board the PTV and integrates it with the ground cellular network, and [9], which discusses a similar approach based on ATM technology. However, it appears that an on-board LAN with readily available support for TCP/IP is more attractive, due in particular to the increasing trend of all-IP based solutions to information access, which allows it to integrate seamlessly with the rest of the Internet. In this article, we therefore focus on TCP/IP-based on-board infrastructure.

bandwidth to the aggregated MWIC as they would to the individual users. This may limit the scale and spectrum of activities that can be supported on-board a PTV.

Furthermore, since a single MWIC is shared among all passengers on the PTV, any changes in the quality of the wireless link (such as fading or temporary outages) simultaneously affect a large number of users. Many factors may cause brief outages in the MWIC, including some that are exacerbated by the nature of typical PTV routes; for example, outages are likely to occur when a train enters a tunnel with no wireless coverage, or a bus passes through a crowded street with high levels of noise and signal interference. If such outages cause the MR to lose and re-establish its connection, the associated overheads (e.g. parameter reconfiguration) are borne by the entire population of on-board users. In cellular connections in particular, outages can also be caused during handoff of the OBLAN between adjacent cells. Furthermore, if the handoff fails due to lack of resources in the target cell, the link may remain down until the PTV reaches and successfully connects to the subsequent cell on the route. Even when the PTV is stationary, outages can still happen due to events such as *intra-zone handoff* or cell selections; empirical evidence suggests that such temporary outages may be common in packet cellular services [12].

An additional effect resulting from the aggregated mobility feature is that of a variable network topology. Since entire IP subnets, rather than individual devices, are mobile and change their point of attachment to the Internet, the Internet topology itself is varying — unlike with personal mobile communications, where movement of a device between networks requires merely a new address binding, but does not cause a change in the network topology itself. This observation requires special support from the Internet protocol stack. The basic support protocol (an extension of Mobile IPv6) proposed by the IETF Network Mobility (NEMO) working group [7] is an important first step in this direction, but much work remains to be done in contexts such as QoS support (e.g. extensions of RSVP), IP multicast, routing, and more.

B. Transparent mobility

In the OBLAN approach, user devices connect only to the OBLAN, which is moving along with the passengers. Therefore, although the passengers are mobile users, there is no relative mobility between passenger devices and on-board infrastructure. The users remain connected to the same IP subnet, and the mobility of the vehicle (and the LAN) remains transparent to the passengers. Along with the above, the wireless link is no longer the last hop on the users' end-to-end connections, therefore the end devices cannot directly monitor the state of the wireless channel. While these observations carry obvious advantages in terms of simplicity of the end user hardware/software, they also mean that many common cross-layer optimizations, relying on the application being notified about the status of the wireless link (e.g. voice or video connections adapting their coding to the available bandwidth and error rate), are no longer possible. Similarly, the end-to-end transport layer, in many cases operating with the TCP protocol, can no longer distinguish between congestion losses and those that are due to signal outages; taking any data loss as a sign of congestion, as TCP commonly assumes, leads to all on-board connections needlessly reverting to an initial ("slow-start") state after any brief outage, which, as described above, may happen frequently in public transport environments.

Since the MWIC connectivity and mobility is entirely managed by the MR on behalf of all users on the OBLAN, whereas application-layer optimizations can only be performed at the endpoints, resolving the mobility transparency issue requires cooperation between the MR and the user devices, which may be either implicit or explicit. Implicitly, there are certain actions that the MR can take to enhance performance of applications without their knowledge, from prioritizing packets (or, for multi-homed OBLANs with several MWIC, choosing the proper wireless interface) depending on the link state and the packet contents, to tampering with well-known fields of common protocols (e.g. the receiver window size in TCP) so as to help the application overcome the effects of link quality variations. Explicitly, the MR may notify the on-board users devices about changes in the MWIC status to allow cross-layer optimizations similar to those prevalent in personal mobile communications. Obviously, this will require standardization of new protocols through which such notifications will take place.

C. Dynamic population

Unlike typical static Internet environments, from enterprises to homes, where users remain connected for extended periods of time and tend to use the same network services repeatedly, public transport OBLANs attract only transit passengers who use the services for short periods of time; for example, with passengers boarding and leaving at bus stops, the LAN user population can become very dynamic. The volatility of the user population raises several difficulties. First, the MR, as well as the wide-area network operator, are unable to estimate the total bandwidth requirement and provision the MWIC accordingly for any significant length of time. Second, the little time that is available to profile the user behavior makes it more difficult to offer personalised services to passengers with diverse background and habits, as well as manage the on-board cache contents efficiently.

D. Repetitive and predictable mobility

The nature of public transport mobility is such that trips are conducted repetitively and according to a pre-determined timetable (with the obvious exceptions of emergencies and operational constraints affecting on-time performance). As a result, the mobility of on-board networks can be predicted with a high degree of accuracy. This observation can be used to mitigate many of the issues raised above. Indeed, most common reasons for major variations of the wireless link quality (including complete connectivity disruptions) can be attributed to location, such as crossing an underground tunnel, being obscured by a high-rise building, or simply leaving a coverage area. Similarly, the user population and their demands, while dynamic and volatile in general, can exhibit repetitive location-dependent patterns; e.g., the demand is likely to be higher in central locations of a city than in outlying suburbs, and regular commuters are likely to pose similar requirements on a daily basis.

The quest for taking full advantage of the predictability of OBLAN mobility presents two kinds of research issues. The first involves designing systems suitable for on-board deployment to measure and record empirical information, and developing algorithms to analyse it so as to discover the necessary patterns and dependencies between environmental parameters (i.e. location, time of day, *etc*), on one hand, and the wireless connection quality and user demands, on the other. The second entails designing enhanced protocols and applications that draw on the mobility predictions and the discovered patterns and dependencies (even though these may not always be fully accurate and complete) to improve the network performance perceived by OBLAN users. For example, the MR (equipped with a location tracking device, such as a Global Positioning System (GPS) receiver) may be able to notify the on-board users in advance about the impending signal loss before entering an underground tunnel, allowing ample time for the applications to adapt to the situation (e.g. perform an orderly clean-up before freezing communications or otherwise switching to an ‘off-line’ mode). More generally, the advance knowledge of the route helps the MR predict and reserve the resources it will need from the wide-area network, thereby reducing the likelihood of disconnections and similar undesirable events when crossing cells or entering congested zones; as a result, applications subject to admission control due to strict resource and QoS requirements (e.g. VoIP calls) can benefit from a more intelligent admission control procedure, that takes into account the anticipated, rather than just the current, availability of resources along the PTV route.

IV. OCEAN Project

The goal of the OCEAN (On-board Communication, Entertainment, And informatioN) research project at the University of New South Wales [8] is to develop efficient techniques for providing on-board mobile access to global information sources, and, in particular, advance the fundamental understanding of the related research issues raised in the previous section. A glance over these issues reveals that research confined solely to the networking discipline, although very important, is not sufficient on its own to realize the goal of scalable and adaptive infrastructure for on-board information access. To that end, research in the area of data and knowledge management is required as much as in networking. Accordingly, in the OCEAN project, we adopt an integrated and synergetic approach that combines research from both these fields. This is particularly evident in the context of exploiting the mobility predictability, where tools from the data management discipline (in particular, data collection and knowledge mining) are used to advance the research of networking problems. Accordingly, we begin this section with a detailed discussion of the approaches we use to exploit the route repetitiveness and predictability. We then provide a sample of other research topics and results achieved in the course of the project.

A. Predicting the connection factors

We now describe the approach used to implement the process of predicting the wireless link quality and the traffic demands and sojourn times of passengers. For brevity, we henceforth refer to these parameters, as well as any additional such information that can help the communication protocols and applications achieve a better service quality, as the OBLAN's *connection factors*. First, one needs to track the relevant environmental parameters — most importantly, location (which can be obtained by GPS or other, possibly proprietary, means), as well as any additional ones that have a potential impact on the connection factors, such as time of day, weather conditions, vehicle velocity, *etc*. Together with measurements of the actual connection parameters (e.g. signal strength, available bandwidth, and user traffic), these are recorded in a historical database, which is eventually used to make the predictions, by comparing the current situation with similar past patterns where changes in the link quality or user demands are known to have occured. Finally, the information coming out of the prediction process — for example, warnings about impending disruptions — are distributed to the on-board user devices, which are expected to use them to make informed decisions and actions at the application, session, and even transport layer.

To back our view that location is indeed the major environmental parameter affecting the connection factors, and particularly the wireless signal quality, in most typical scenarios, we conducted a detailed study in [13] of the cellular network signal strength under a variety of conditions in a metropolitan area. Specifically, we recorded the signal strength of Vodafone's GPRS network in different locations in Sydney, Australia, focusing in particular on the major train and bus routes.[†] Our study found that the signal strength varies significantly among the different locations, whereas the variation is much smaller for a given location across different times. This allows us to conclude, at least for cellular communications, that even a simple predictor of wireless signal quality variations, based solely on the vehicle's location, is likely to be adequate for most applications.

The precision with which the PTV location needs to be gauged may differ with the particular types of transport and environments. A typical GPS receiver can achieve a precision in the order of a few metres; thus, e.g., if it is used to detect an imminent tunnel crossing on a fast long-distance train that travels at a velocity of tens of metres per second, the uncertainty of the location measurement extends the warning period by a fraction of a second at worst. On the other hand, a higher precision may be required in a metropolitan setting, where the dependence of the radio signal quality on the location is more erratic and vehicle velocities are lower. This higher precision may be achieved by either publicly available (e.g. cellular triangulation) or proprietary methods (e.g. sensors mounted on rail tracks). We point out that, to this end, we are encouraged by the recent advances and rising interest in vehicle collision avoidance systems, which use visual, ultrasonic, and other sensors to detect the distance from the vehicle in front and from the sides of the road (e.g. [14], [15]). Clearly, once the underlying technologies for such systems become commonplace, they can also be used to achieve reliable and precise location measurements to feed the connection factor prediction engine.

At the heart of the prediction process lies the analysis of the historical records of the connection factors, obtained during the recurring PTV trips. In principle, such an analysis is conducted using classification methods from the established fields of data mining or neural networks, where a classifier trained with the historical data is capable of assessing the connection factor value (e.g., in the simplest version, determining whether the wireless link is "down" or "up"), given the value of the location and other environmental parameters. However, prediction brings about several unique challenges that make it quite different from most classification problems studied in the literature [16]:

- **Anticipated classification.** Prediction requires the classifier to operate on *future* anticipated values of the environmental factors (e.g. location), rather than present ones. Calculating the anticipated location a certain time into the future requires the consideration of a multitude of secondary factors (present vehicle velocity, timetable, *etc*), which affect the confidence level of the prediction outcome. The situation is further complicated by the fact that different applications may require different advance warning requirements, even if running simultaneously — e.g., for a warning about an impending link outage to be useful, it generally needs to be available to the application sufficiently in advance to allow at least one round-trip exchange with the remote party, which hence depends on the latency of the network connection.
- **Data streams.** In most cases, it can be expected that the required advance warning period will be measured in seconds; hence, the measurements on which the classification algorithm operates must be collected with a time granularity no coarser than that. Consequently, each vehicle will collect a huge stream of data during every single trip. Although the data can be used to train the classification and/or prediction model, it may not always be necessary to store all of it in the historical database. Selective data storage schemes, which attempt to reduce the data storage overhead while preserving as much information as needed for the data analysis and

[†]We note that, undoubtedly, similar measurement studies are performed internally by the cellular service providers. However, to our best knowledge, their results are not publicly available.

classification, are likely to prove valuable in this context; in particular, recently proposed ideas in *streaming data mining*, based on the concept that more recent data are stored in more detail [17], are particularly useful in this regard.

- **Distributed event records.** In practice, classification and prediction models cannot achieve full efficiency if they are only trained with local data collected on each individual PTV, since any particular vehicle might operate on different routes from time to time. On the other hand, any PTV is nearly always a member of a large fleet that travels on the same route many times per day, and collectively possesses a large volume of fresher raw data. The data collected by multiple vehicles may be combined for usage in prediction model training. However, it is unfeasible to maintain a centralized database with full real-time sharing of the collected measurements, due to the huge bandwidth it would require. Yet, even a limited, low-bandwidth cooperation may be able to bring about significant improvements in the success rate of the prediction process; e.g., a background broadcast of low-volume "digests" of irregular events, such as detours from the normal route due to road construction or accidents, may help subsequent vehicles to avoid making predictions based on stale records. A more effective alternative is to combine multiple classifiers learned from different event logs or combine collaborative filtering with classification techniques [18], [19]. This is also important for passenger profiling, as it allows profiling information to be re-used when passengers transfer between PTVs during the trip.

B. Anticipatory on-board network management

The ultimate goal of the disruption prediction process is to enhance the performance of network applications and protocols. We now proceed to present a variety of enhancements, spanning all layers of the network protocol stack and data management procedures, which demonstrate the benefits from advance knowledge (even if only partial and imprecise) of the wireless link quality and the on-board user population and demands.

- **Freezing transport-layer connections.** One of the simplest ways to take advantage of advance knowledge of an imminent wireless disruption is to employ a variation of the Freeze-TCP protocol [20]. Freeze-TCP was originally designed to prevent undesirable timeouts (which lead to unnecessary slow-start and congestion avoidance phases of an already active connection) when a mobile host is handed off to a new wireless base station. To achieve that, once notified of the impending handoff, the mobile host's TCP implementation transmits an acknowledgment carrying a zero window advertisement (ZWA), causing the remote (fixed) host to freeze any further data transmission and enter a probing mode, during which it periodically attempts to send one-byte probes until the mobile host is reconnected. In addition, as soon as the mobile host is reconnected at the new base station, it transmits a triple acknowledgment with a non-zero window advertisement (NZWA), enabling the remote host to resume its data transmission.
  The Freeze-TCP approach can readily be implemented in an OBLAN context, with the exception that ZWA and NZWA messages are generated by the MR transparently on behalf of the end-user devices, whenever an imminent MWIC disruption is predicted (and after the disruption is over, respectively). In [21], we developed a Markov model of Freeze-TCP in a setting of a wireless connection prone to outages (i.e. alternating between 'down' and 'up' states) and a prediction mechanism that is only able to predict a portion of the outages. We studied the dependence of the protocol performance on prediction quality parameters (e.g. percentage of correctly predicted outages), and found that investing in a good prediction technology, which is able to predict a high percentage of outages in a timely manner, becomes increasingly important as outages occur more frequently and/or TCP connections use larger windows.
- **Multihoming.** As described above, one of the advantages of the OBLAN architecture is the ability of a MR to be multihomed, i.e. connected through several wireless interfaces, typically involving different technologies or different service providers. This allows the on-board users to be continuously connected even when one of the interfaces is disrupted for a long time, e.g. in a scenario where a long-distance train leaves the coverage area of a metropolitan cellular network and continues to maintain communications via satellite. In order to allow the MWIC transition to be smooth and transparent to the users, the disruption of the existing link must be predicted well in advance, since the vertical handoff procedure normally entails a significant overhead of several round-trip exchanges. Furthermore, the management of several wireless interfaces becomes quite involved if they can be used in parallel; then, handovers may be carried out not only when a MWIC is disrupted, but even

in the course of normal operation, e.g. to rebalance the load after several users terminate their sessions on one of the links. Indeed, the problems arising in the management of multiple-MWIC networks are an active topic in the IETF NEMO working group [22].

In [23], we described a multihoming model based on generic volume charges for traffic in the different connections and pose the traffic distribution problem as a profit maximization problem for the on-board network operator. We study both an *in-transit* and a *pre-transit* management scheme, where the latter is assumed to have advance knowledge of changes in the traffic demands and resource availability. We find that, even though this prior knowledge does not result in any additional profit, it has a direct effect on the level of service disruption for the on-board users in the optimal solution.

- **Admission control and resource reservation.** The aggregated mobility feature of the OBLAN means that it presents itself to the MWIC provider as a single user with an unusually large bandwidth requirement. Unfortunately, in most access technologies (e.g. cellular networks), larger resource demand translates to an increased blocking probability. This observation may require a special approach on the provider's behalf to managing the bandwidth resource so as to provide a reasonable quality of service both to the MR (and the aggregated on-board users) and to individual users elsewhere. In [24], we study and compare two methods of preferential treatment of MRs in cellular networks: in one, a portion of capacity is reserved for exclusive use by MRs, using feedback control tools to achieve a desired call blocking probability; in another, no capacity is set aside for MRs but lower-priority (i.e. non-MR) calls are interrupted if necessary to accommodate new (or handed-over) MR calls, and admission control is applied to keep the call interruption probability below a desired level. Our results indicate that both techniques are viable under realistic settings and are able to achieve reasonable blocking and interruption probabilities for all users involved.

  Moreover, it is clear that the admission control performance can be improved even further by reserving bandwidth in advance, taking advantage of the predictable vehicle route. The concept of using advance reservation (aka "booking ahead") of resources has received some attention in fixed networks (see, e.g., [25] and references therein). In [26], we applied this concept to cellular networks serving applications with predictable resource requirements (e.g. call start time, duration, and mobility), as is the case in the OBLAN architecture. To offer a resource booking facility for arbitrary start times and durations, the network must store, process, and access large volume of configuration data that describes the future allocation of critical network resources. In particular, the admission control functions must have fast and scalable access to these configuration data for making effective decisions regarding the acceptance and rejection of both immediate and future service requests. We proposed an architecture for a distributed reservation database architecture that allows an efficient and scalable access to the advance reservation data. Our experiments demonstrated that this architecture, which features proactive processing and delivery of resource configuration data to the admission control functions in the cellular base stations, can achieve admission control response times that are faster by orders of magnitude than traditional, query-based mechanisms.

  In addition, the aggregated mobility of OBLANs is also helpful for resource reservation in the Internet beyond the MWIC itself, since it means that resource requests (using protocols such as RSVP) can be aggregated to save overhead. In [27], we propose *On-Board RSVP*, an extension of the Internet resource reservation protocol, where resource reservation requests from on-board applications are accumulated in a queue for a while before being transmitted in a single aggregated packet over the tunnel from the MR to its home agent. We further study alternative aggregation policies in [28] and compare the performance of policies where the aggregation threshold is based, respectively, on time, number of requests, and total resource requirement in the queue. We show that, in general, aggregation is able to achieve significant savings of overhead, with negligible impact on the performance of individual applications, and that the time-based policy is the most cost-efficient one under a variety of operating conditions.
- **Location-based services.** A location-based service (LBS) is a kind of personalized service where the service is dependent on positional information [29]. An OBLAN is an ideal platform for providing LBS, due to the predictable mobility of the PTV and the fact that the PTV location can be accurately determined (which is, indeed, an important part in the prediction process). Moreover, the aggregated mobility of the OBLAN means that the overhead of location tracking, a key component of most LBS approaches, is greatly reduced. There are two kinds of LBS that can be provided:

- *General LBS*. Many LBS require only the location information of the PTV without considering passengers' personal preferences. Examples include estimation of arrival times, maps of current location, and location-aware advertising (e.g., e-coupons). This kind of location-aware information can be cached by the on-board data server and made accessible to the passengers without any prior collection of personal information.
- *Personalized LBS*. An example of a passenger-personalized LBS is a PTV reminder service, which informs a passenger where to alight once the target destination of that passenger is registered. For regular travellers, the PTV reminder can automatically identify the target stop based on their past travel history, and even provide estimated arrival times of connecting PTVs.

The key research question in the provision of LBS related to managing the content in the on-board server. Clearly, prefetching LBS data (such as maps and arrival times) into the on-board server reduces the response time and the wireless communication cost for each individual user, particularly since the wireless connection cannot be guaranteed to be always available when the user may need it. On the other hand, since some of the data (e.g. e-coupons) may only be valid for a limited time until expiring, keeping a large amount of pre-fetched data can become too costly. Therefore, efficient management of the on-board server content is greatly assisted by the prediction of the wireless connection quality and user demand, based on past pattern analysis.

It is important to note that, in all of the enhancements suggested above, the MR can act on the predicted information itself, i.e. transparently to the applications. Thus, such enhancements can readily be implemented in future mobile routers, without necessitating any upgrades or modifications at the user end. However, we expect that as on-board mobile services based on the OBLAN architecture gain in popularity, future user applications running in the end devices will be designed from the outset to withstand frequent variations of the connection factors, and to cooperate with the MR by actively receiving prediction notifications.[†] In particular, such applications will be able to react to changes in the predicted link quality, e.g. by freezing, terminating gracefully, or switching to an application-specific 'offline' mode upon receiving a warning about an imminent disruption. Moreover, applications may even be able to proactively request medium-term 'forecasts' of the wireless connection quality expected during the trip so as to tune their actions, e.g. to refrain from initiating an important transaction just before the link is likely to go down, or to adapt the transmitted content in anticipation of future disruptions (for instance, a variable-rate streaming multimedia application could plan its coding rate strategy in advance). For any of the above to be achievable, clearly, it is crucial that protocols are developed and standardized for notifications and predictions to flow between the MR and the end devices; in this regard, a major challenge lies in designing flexible data formats that can convey a variety of information, from simple warnings about imminent outages to more convoluted forecasts that provide 'soft', detailed probabilistic information about anticipated connection states and confidence levels. This area remains an exciting and important direction for future research.

C. Other research issues

While much of our research revolves around the key feature of mobility repetitiveness and predictability, there are a number of other issues requiring attention due to the OBLAN architecture. The remainder of this section provides a sample overview of such issues tackled in the OCEAN research project and the results obtained.

- **Security.** The OBLAN environment requires some mobile communications infrastructure — namely, the mobile router — to be co-located in the vehicles, in the immediate vicinity of the users. The lack of the same level of physical protection typically enjoyed by stationary routers located within secure buildings, coupled with the fact that all communications are conducted over a wireless medium, makes MRs more vulnerable to various security threats, including attacks that are impossible or very hard to accomplish with stationary routers. Consequently, protocols to support OBLAN connectivity must be designed with such vulnerabilities in mind, and in particular, be able to protect the OBLAN users as well as the wide-area network infrastructure (including other OBLANs) from the possible impact of compromised mobile routers.

  With the IETF Network Mobility (NEMO) protocol [7] emerging as a standard for providing the basic mobility support for OBLANs, it is important to consider its particular security vulnerabilities. The NEMO protocol is largely an extention of Mobile IPv6, the security of which has been studied thoroughly (see, e.g., [31]).

[†]The need for novel applications and protocols able to cope with frequent link disruptions has been gaining increasing attention in the research area of *delay-* (or *disruption-*) *tolerant networking* (DTN), that is the focus of the IRTF DTN Research Group (DTNRG) [30].

However, its use requires the configuration of a number of settings in the MR before user data can be transported through it. While these settings can be permanently configured before the MR is installed on a vehicle, it may be preferable to dynamically configure as many of these settings as possible to allow for more dynamic updates in networks and services. The Dynamic Home Agent Address Discovery (DHAAD) protocol allows a MR to dynamically discover a new HA on its home network, and may be used routinely by a MR whenever it loses connectivity with its home network, due to prolonged outage in the MWIC or other problems.

In [32], we have performed a security analysis of the DHAAD protocol and pointed out the fact that DHAAD requests, being unauthenticated, can be easily sniffed over the wireless link by an attacker. The attacker can then reply to the MR with an invalid HA address list, thus preventing it from establishing a tunnel with its home network, resulting in a denial of service to the on-board users. This simple example highlights the need for a thorough study of potential security holes in protocols that are likely to be used by a mobile router. It is important to highlight that the solution provided in [32] for securing the DHAAD protocol is fast (since the MR needs to re-establish a tunnel to a HA immediately after an outage) and light-weight in terms of bandwidth consumption (since wide-area wireless bandwidth is scarce and expensive).

- **Summarisation and query processing.** The inherent features of the on-board environment require intelligent support for personalized and timely delivery of information to users, with one of the key considerations being reduction of the wireless communication costs. Consider, for instance, a scenario of web-shopping, where a user must browse a multitude of websites offering a certain desired product before eventually making a choice. To reduce the user's effort as well as the wireless communication cost, and to make the information access process more efficient, there is a need to provide a smart, integrated query processing in the on-board architecture.

  We propose an integration framework to efficiently integrate query information portals and services [33], [34]. Our approach is based on a hybrid of a peer-to-peer data sharing paradigm and a Web service architecture. Each information portal or service is known as an e-catalog. E-catalogs catering for similar customer needs are grouped and registered as members of a *catalog community*. Catalog communities serve as domain-specific mediators and are linked as *peers*. Peer links among communities are established based on the similarity/relevancy of domains they represented. Queries are initially received and processed by a single community. If the community cannot satisfy all the requirements specified in a query, the query is divided into several sub-queries and each of the sub-query is routed to relevant peers who are able to answer the sub-query. In each community, a set of members is selected to answer the sub-query received. The community produces the final result for the query with a join operation of all sub-query answers.

  As queries are answered by a combination of related information sources, the information integration framework significantly reduces passengers' efforts in locating information of interest, as well as wireless communication cost. However, when the number of information portals and services grows large in communities, it is time-consuming and costly for the mediator to effectively identify the set of members to answer a query. Efficient techniques to identify a subset of members that are most likely to provide relevant answers to a given query are therefore the key to make our framework scalable. Building community summaries can be an effective way to perform content-based selection of e-catalog members, and the questions of how to summarize the e-catalog member's database, keep the summaries current and updated, and use them to select members, are all being addressed in ongoing research.

V. Conclusion

The vision of ubiquitous communications and information access for mobile users, and particularly for passengers on public transport, requires novel communication solutions to be deployed on board of public transport vehicles. This article has considered the *on-board LAN* architecture, which is gaining increasing popularity and already becoming deployed in some initial commercial offerings. We presented a detailed discussion of the characteristical features and some of the research challenges that must be addressed for the architecture to achieve its full potential and facilitate a widespread deployment of scalable and flexible support for ubiquitous connectivity on public transport.

ACKNOWLEDGMENTS

The OCEAN project is supported by Australian Research Council Discovery Grant DP0452942. The authors wish to thank the many students and staff who contributed to the OCEAN project.

REFERENCES

[1] Nautilus 6 working group: http://www.nautilus6.org.
[2] icomera: http://www.icomera.com.
[3] PointShot Wireless: http://www.pointshotwireless.com.
[4] 21Net: http://www.21net.com.
[5] Inmarsat's Swift64: http://www.inmarsat.com/swift64.
[6] Connexion by Boeing: http://www.connexionbyboeing.com.
[7] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert. RFC 3963: Network mobility (NEMO) basic support protocol, January 2005.
[8] Project OCEAN — On-board Communication, Entertainment and informatioN: http://ocean.cse.unsw.edu.au.
[9] K-D. Lin and J-F. Chang. Communications and entertainment onboard a high-speed public transport system. *IEEE Wireless Communications*, 9(1):84–89, February 2002.
[10] J. Irvine, J-P. Couvy, F. Graziosi, J. Laurila, G. Mossakowski, and P. Robin. System architecture for the MOSTRAIN project (mobile services for high speed trains). In *Proc. IEEE Vehicular Technology Conference*, pages 1917–1921, Phoenix, AZ, May 1997.
[11] M. Uhlirz. Concept of a GSM-based communication system for high-speed trains. In *Proc. IEEE Vehicular Technology Conference*, pages 1130–1134, Stockholm, Sweden, June 1994.
[12] R. Chakravorty, S. Katti, J. Crowcroft, and I. Pratt. Flow aggregation for enhanced TCP over wide-area wireless. In *Proc. IEEE Infocom*, pages 1754–1764, San Francisco, CA, March 2003.
[13] I. Chan, A. Chung, M. Hassan, K.-C. Lan, and L. Libman. Understanding the effect of environmental factors on link quality for on-board communications. In *Proc. IEEE Vehicular Technology Conference (VTC)*, Dallas, TX, September 2005.
[14] E. Dagan, O. Mano, G.P. Stein, and A. Shashua. Forward collision warning with a single camera. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 37–42, Parma, Italy, June 2004.
[15] K-T. Song, C-H. Chen, and C-H.C. Huang. Design and experimental study of an ultrasonic sensor system for lateral collision avoidance at low speeds. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 647–652, Parma, Italy, June 2004.
[16] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, August 2000.
[17] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu. A framework for clustering evolving data streams. In *Proc. International Conference on Very Large Data Bases (VLDB)*, pages 81–92, Berlin, Germany, September 2003.
[18] S. Zhang, I. Cohen, M. Goldszmidt, J. Symons, and A. Fox. Ensembles of models for automated diagnosis of system performance problems. In *Proc. International Conference on Dependable Systems and Networks (DSN)*, pages 644–653, Yokohama, Japan, June 2005.
[19] J. Wang, A.P. de Vries, and M.J.T. Reinders. A user-item relevance model for log-based collaborative filtering. In *Proc. European Conference on Information Retrieval (ECIR)*, London, UK, April 2006.
[20] T. Goff, J. Moronski, D.S. Phatak, and V. Gupta. Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments. In *Proc. IEEE Infocom*, pages 1537–1545, Tel-Aviv, Israel, March 2000.
[21] A. Baig, M. Hassan, and L. Libman. Prediction-based recovery from link outages in on-board mobile communication networks. In *Proc. IEEE Globecom*, Dallas, TX, November 2004.
[22] C. Ng, E. Paik, and T. Ernst. Analysis of multihoming in network mobility support, February 2005. Internet draft: draft-ietf-nemo-multihoming-issues-02.txt (work in progress).
[23] A. Chung and M. Hassan. Optimizing profit and performance for multi-homed mobile hotspots. In *Proc. IEEE ICC*, Seoul, Korea, May 2005.
[24] M. Hossain and M. Hassan. Regulating call blocking of mobile routers in wireless cellular networks. In *Proc. IEEE Vehicular Technology Conference (VTC)*, Dallas, TX, September 2005.
[25] L. Lewin-Eytan, J. Naor, and A. Orda. Admission control in networks with advance reservations. *Algorithmica*, 40(4):293–304, September 2004.
[26] A. Sun, M. Hassan, M.B. Hassan, P. Pham, and B. Benatallah. Fast and scalable access to advance resource reservation data in future cellular networks. In *Proc. IEEE ICC*, Istanbul, Turkey, June 2006.
[27] M.A. Malik, S.S. Kanhere, M. Hassan, and B. Benatallah. On-board RSVP: An extension of RSVP to support real-time services in on-board IP networks. *Springer Lecture Notes in Computer Science (LNCS)*, 3326/2004:264–275.
[28] M.A. Malik, S.S. Kanhere, and M. Hassan. Aggregation policies over RSVP tunnels. In *Proc. IEEE Vehicular Technology Conference (VTC)*, Dallas, TX, September 2005.
[29] C. Steinfield. The development of location based services in mobile commerce. In B. Preissl, H. Bouwman, and C. Steinfield, editors, *E-Life after the Dot Com Bust*. Springer, 2004.
[30] DTN Research Group: http://www.dtnrg.org.
[31] T. Aura. Mobile IPv6 security. *Springer Lecture Notes in Computer Science (LNCS)*, 2845/2003:215–234.
[32] S. Qian, M. Lei, H. Petander, K.-C. Lan, and M. Hassan. On securing dynamic home agent address discovery of on-board mobile router in mobile IPv6 networks. In *Proc. International Conference on Telecommunications (ICT)*, Capetown, South Africa, May 2005.
[33] H.-Y. Paik, B. Benatallah, and F. Toumani. Towards self-organizing service communities. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 35(3):408–419, May 2005.
[34] H.-Y. Paik, N. Mouaddib, B. Benatallah, F. Toumani, and M. Hassan. Building and querying e-catalog networks using P2P and data summarisation techniques. *Journal of Intelligent Information Systems*. To appear.