Context-Aware Resource Management for Mobile Servers

Christopher Ververidis, Elias C. Efstathiou, Sergios Soursos, George C. Polyzos

Mobile Multimedia Laboratory,
Department of Informatics,
Athens University of Economics and Business,
10434 Athens, Greece

{chris, efstath, sns, polyzos}@aueb.gr
http://mm.aueb.gr/

Tel.: +30 210 8203 693, Fax: +30 210 8203 686

Abstract
Mobile operators wishing to deploy higher-level value-added services face various costs attributable to the introduction of new software and hardware components into their networks. Considering that today's wireless devices are becoming increasingly capable of running complex applications with server-like features, we could envision the case where mobile operators "outsource" value-added services provision to subscribers equipped with such devices, which will thus be acting as mobile servers. Services may either be attributed to these mobile servers by the mobile operator or be created in an ad-hoc manner by the subscribers and could range from file sharing applications to location based, context-aware service offerings, such as real-time road traffic information. Network operators benefit from this scheme not only because they enrich their offerings at minimal cost, but also because traffic towards the outside of their networks (to obtain similar services) is limited. Clients benefit because these "internal" services are accessed from within the core network of the provider at high-speed since external and potentially congested routes are avoided.

1. Introduction

Considering the convergence of computing and telecommunications technologies and with the tremendous success of the Internet, the World Wide Web, and Mobile Communications, the next step is expected to be the *Mobile Internet*. The main promise of the Mobile Internet is to satisfy user needs for anywhere/anytime wireless access to information and services, including Location Based Services (LBS). The vision of cost-effective, fast and ubiquitous access to the Mobile Internet is becoming a reality through the introduction of new access technologies.

The emergence of the aforementioned low-cost wireless access technologies – with IEEE 802.11 WLANs being the most prevalent – reduces infrastructure costs, thus blurring the line between traditional (large-scale) providers and consumers of digital goods. The Peer-to-Peer (P2P) paradigm, which is this decade's version of the Internet's "end-to-end" argument, involves the exploitation of idle resources and the provision of goods by numerous mini-providers. Today, with the introduction of WLAN technologies, we have abundant access bandwidth, storage, and processing power at the edges of the network. There, these resources may very well be under the control of individuals or small organizations that may use them to provide customized higher-level services as they see fit.

Today's Personal Digital Assistants (PDAs) and Mobile Phones are converging towards, so called, "next-generation" terminals, which combine traditional voice capabilities with enough processing power to handle generic computational tasks, usually coupled with high-resolution colour screens and easier ways to obtain user input (voice recognition will be a significant technology that will enable convenient user control over small devices). Built-in cameras for taking still pictures or videos are becoming increasingly common. On the operating systems front there is still no one standard but, on the other hand, middleware technologies such as Sun's Java Micro-Edition and Microsoft's .NET Compact Framework will allow developers to target their applications at an increasing number of these devices. At the very least, we can always assume that the IP suite of (v4 or v6) protocols will be supported on top of a device's wireless access interfaces. Packet-based technology will allow efficient file-transfers, streaming media and interactive communications at several Quality-of-Service (QoS) levels.

The objective of mini-cellular providers will be to maximize the value their wireless cells offer to end users, perhaps by making some wireless cells more "attractive" than others. The problem of allocating resources in a dynamic way to the users that value them most is non-trivial. Defining and achieving cell throughput stability in the context of services being provided by mobile servers is one of the objectives of this paper.

The structure of the remainder of this paper is as follows. In section 2, we introduce a new business model for Location-Based Services. In section 3, we define the principles of operation of the service we propose and we present an example. In the following section, we describe the required network architecture. Next, we describe how the proposed architecture can be implemented. Finally, we discuss some open issues and present our conclusions.

2. LBS Deployment – A new Business Model

A recent market research [1] showed that mobile subscribers would even consider giving up their old mobile network operator in favour of one that would provide them access to LBS. The same research also indicated that subscribers would be willing to pay more as a monthly fee for these services. Moreover, mobile operators, having already established large customer bases, seek new ways to ensure customer loyalty by offering new types of services. Location Based Services are identified as the most promising ones since they provide:

- Innovative, useful services, which attract new customers and enhances customer loyalty to the provider
- Revenue increase due to traffic generated by the use of such services.

In the past, LBS provision [2] was based on three generic business models. Their distinguishing characteristic was the level of involvement the carrier had in providing these services; the carrier faced bigger costs but got the full revenue; with the partners, the value chain would be extended, but the generated revenue would have to be shared.

In this paper, we propose a new business model where the mobile subscribers themselves are assuming the role of providing LBS and more generally, context-based services to other mobile subscribers in a P2P manner. We assume that the subscribers

achieve that at no significant cost, since providing the service comes as a by-product of their mobility. The carrier only offers network connectivity and provides the mobile subscribers with the needed positioning infrastructure. In our business model, the carrier manages to keep network transactions relating to LBS content within its network. This means that the number of contracts with external content providers can be reduced, alongside the related traffic that would have crossed network boundaries. At the same time, by leveraging the storage space provided in next generation mobile terminals, its own storage needs and the associated Operation, Administration and Maintenance (OA&M) costs are kept at a minimum. Thus, overall carrier costs are reduced, while user satisfaction is increased (positive externalities). By letting end-users offer this advanced service through their own terminals, other indirect benefits can also be achieved (such as community-building spirit and increased customer satisfaction). Except from keeping costs down, the carrier may also increase its revenues since all mobile subscribers interested in location-based services are willing to pay a little more for obtaining them. Users acting as mobile servers can also benefit directly, assuming a brokerage scheme (with the carrier acting as the broker) that would direct funds from clients to servers.

We assume that a mobile network operator (deploying either 3G or WLAN networks, or both) adopts this new business model in order to provide its mobile subscribers with the ability to offer cell-local services, i.e. providing content that is location (and time) sensitive.

3. Service Provision Environment

Consider the case that along side the central avenues of a city there exist several Access Points (in the case of WLAN) or Base Stations (in the case of UMTS). We will use the term Access Point – APs – to refer to both. A single AP provides coverage in a specific part of the road, e.g. a frequently congested part and its whereabouts. In this coverage area, the network operator advertises the provision of free-of-charge photo-series that depict the situation of the road at a specific time. Such photo-series (or even low quality/length videos) may help the drivers passing by to decide whether they want to take the specific avenue or to change their course due to the heavy traffic ahead. To do so, the network operator would have to install its own technological equipment and store such information, or cooperate with an external service provider that possesses such technology.

By applying however the business model we outlined before, we give the network operator another alternative. He can let the mobile subscribers provide such a service and exchange, with other subscribers, photos or videos that depict the traffic conditions that they had encountered till that moment. Thus, the operator relies on the existence of mobile subscribers, whose devices have the ability to capture pictures (and videos), and who can therefore act as mobile servers for this task.

These photos may be of no interest outside the AP's coverage area, since they are out of context. The reason is that when the mobile subscriber is outside the AP's coverage (and inside the coverage of another AP) the information that it holds is, outdated and does not apply to the current cell's traffic situation. So, we assume that it is reasonable for the mobile servers to be offering these photos only when inside the aforementioned coverage area. Moreover, if the mobile subscribers that are acting as clients leave the

specific coverage area then this service would no longer be available. This is also reasonable since leaving the specific coverage area may result in increasing the costs for maintaining the file transfer alive. To be more precise, leaving the area of interest may imply leaving the coverage area of a specific WLAN or UMTS access provider. Thus, in order to keep the connection uninterrupted, the packets may have to travel through networks that offer services of lower quality (e.g. GPRS, GSM) and consequently result in higher transmission costs.

One of the main assumptions of our architecture is that the network operator has in place the infrastructure needed for service advertisement and discovery. One solution to the service discovery problem is that the network operator maintains, at the very least, a central database where all mobile servers advertise their services and the mobile clients search in order to find the one that interests them (much like the old centralized-index Napster model). Another approach is to preserve a distributed hashtable-like search mechanism (using a protocol like Chord, CAN or Tapestry for locating servers that possess a specific piece of content). In both approaches, mobile subscribers search inside the coverage area to find mobile servers that offer specific content. The second approach does not add extra burden to the network operator for maintaining a central mechanism (apart from an agent to assist with the distributed hashtable bootstrap protocols). In yet another model, it may not even be required for subscriber to issue a request on his own: the operator can provide him with an index of all the available content when the subscriber enters the specific coverage area.

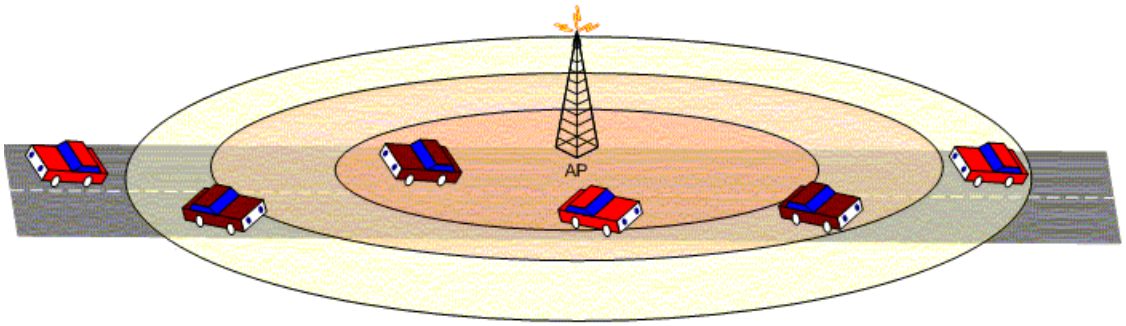
As we mentioned previously, the main goal of this architecture is to provide location-based services to the mobile subscribers of a network operator without the operator having to store local content or having to maintain contracts with external content providers. In order to achieve this, some of the mobile subscribers may offer some content that may be useful for other subscribers too. An important issue here is that the network operator must favour the content transactions that tend to preserve the content inside the coverage area.

Our analysis focuses on two main points: the need of mobile servers for different service provision quality according to their location in the coverage area and the need for maintaining the desired content in the coverage area. In the following section, we introduce the proposed architecture that will support the ultimate goal of the described environment. Note that for simplicity, we assume that there is only one AP providing all the needed coverage. Our model is easily extended to include more than one AP.

4. Architecture

One of the main issues is to provide a metric for estimating the precise or relative position of the mobile servers in the coverage area. Current mobile positioning technologies vary from the simple “Cell Of Origin (COO)” to much more complicated technologies, like the “Observed Time Difference Of Arrival-Idle Period DownLink (OTDOA-IDL)” and the “Global Positioning System (GPS)”. In our approach we assume that one of the simpler approaches is being used (since we don’t want to introduce complicated technologies to the operator’s network). It is reasonable though that if an operator possesses a more precise technology, he may use it instead. Our architecture, therefore, does not rely on GPS-type location precision.

Given the aforementioned positioning model, we divide the coverage area into three concentric circles in order to provide us with a metric for estimating the relevant position of the mobile devices with respect to the AP. The specific bounds of each concentric circle are defined in terms of the signal to interference and noise ratio (SINR), a metric provided by the power readings from the AP. In fact, we introduce four different SINR thresholds in order to define whether a mobile device is located respectively in the first, second or third concentric circle, or if it is out of the coverage area. Note that the use of SINR may mislead the system about the actual position of the mobile subscribers' devices, since interference may exist in the radio channels. An obstacle between the subscriber and the AP may provide the AP with the inaccurate information that the mobile subscriber is more distant. Statistically, however, we assume that the SINR method is good enough.



In our model, according to the position of the mobile server, the operator must provide him with different uplink throughput, in order to assist the server in finishing all the pending transactions before leaving the coverage area. So, when the mobile server is moving outwards, i.e. it is leaving the coverage area, then, more uplink throughput must be allocated to it in order to serve as many requests as it can before going out of reach. These actions are taken by the appropriate network resource management module of each technology. In UMTS it is the Radio Resource Management (RRM) entity located at the Radio Network Controller (RNC).

Since the IEEE 802.11 standards do not attempt to define an entity with similar characteristics to the RNC (an access point controller, essentially), the required entity can be custom-build with standard PC hardware. Assuming a PC equipped with a WLAN card that supports *access point* mode, the PC can be programmed to take over normal access point functions (relaying of frames and/or bridging to the wired network when required) but with an advanced queuing discipline (not just FIFO) which would prioritize according to the cell scheme we presented and drop (or rearrange) any excess packets. Assuming TCP at the senders, the senders would lower their rate for every packet loss. In effect, by using multiple queue service rates and relying on TCP's flow control, this software access point can regulate the uplink rate of every sender.

For example, let us assume that there is a mobile server inside the inner circle and there is also a request from a mobile device that is also located inside the inner circle. If we define that the throughput per flow given to mobile servers in each circle is 10, 20 and 50kbps respectively, then in the aforementioned example the mobile server transmits the photo with 10kbps. Once the mobile server moves to the next circle the transmitting rate increases 20kbps and once the mobile server reaches the outer circle it transmits

with 50kbps. Note that all the above happen if the mobile device requesting the photo is located at the inner circle.

But the mobile client is also moving within the coverage area. It is logical to assume that when a mobile client is located at the outer circle, it is either leaving or entering the coverage area. In the former case, we assume that the transfer is completed. But in the latter case, the mobile client must receive the photo-series (or the video) as quickly as possible in order to decide what to do. So the mobile server must adjust its transmitting rate accordingly. Below we provide a table of (example) actions taken by the network operator with respect to server's transmitting rate in order for the mobile server to serve the incoming requests.

Mobile Server → Mobile Client ↓	Inner Circle	Middle Circle	Outer Circle
Inner Circle	10kbps	10kbps	20kbps
Middle Circle	10kbps	20kbps	50kbps
Outer Circle	20kbps	20kbps	50kbps

When the mobile client is at the outer circle, it is important that he receives the photo-series quickly. When the mobile server is also at the outer circle then the highest possible rate must be assigned to the particular transmission, since the mobile server is leaving the coverage area. As the client moves toward the inner circle, the transmission rate drops since the information provided by the photos will be less useful. Furthermore, as the mobile server moves outwards, the transmission rates increase since the photo must be given a chance to remain within the coverage area.

Another important issue for the network operator is how to provide a mechanism so that the content is kept within the coverage area for as long as possible, so that there is some content to be exchanged between the mobile subscribers and not wait for a new server-enabled mobile subscriber to enter the coverage area.

To achieve this, the file-transfer sessions with clients that are themselves registered as mobile servers must be prioritised, especially if the transmitting server is leaving the coverage area. This would generally assist in keeping the content inside the area of interest. If a mobile server requests a photo from another mobile server which is leaving the area, then this exchange must receive enough transmitting rate so that, at the same time, the requesting mobile server can also serve clients that request the specific photograph, even if the mobile server that initially took the photograph is out of reach.

5. Implementation

It is clear that the actions made by the network operator in order to support the architecture mentioned in the previous section assume the existence of a context-aware QoS scheme taken into consideration by the network management module. Unlike existing QoS schemes for wired and wireless networks, the QoS scheme presented here does not apply to all the outgoing traffic of a mobile server but only affects the transfer of a specific file between two mobile subscribers, in the context of an LBS session. Thus, the outgoing flows of a mobile server may receive different treatment by the network management module, relative to the position of the mobile server, the position

of the mobile client and the type of the mobile client (which may also be registered as a mobile server).

Two different approaches can provide the needed QoS architecture for the proposed content provision environment. Borrowing relevant terms from the IETF protocols, we label one LBS-IS (borrowing the term from IETF Integrated Services [7] and the respective flow-level granularity), and LBS-DF (from IETF Differentiated Services [6] and the respective traffic class-level granularity).

LBS-DF approach:

In this approach, we define three independent QoS classes (1 to 3, with class 3 being the class with the higher priority), corresponding to different transmission rates. In fact, in the AP there exist three different flow-serving queues, each one with a different serving rate. Through the use of a scheduling algorithm (e.g. Weighted Fair Queuing - WFQ) the AP processes the packets arriving to each one of these outgoing queues. According to the position of the mobile server and client, the mobile server's outgoing traffic is placed in the appropriate queue. Thus, if we match QoS classes to our previous table, we have:

Mobile Server → Mobile Client ↓	Inner Circle	Middle Circle	Outer Circle
Inner Circle	QoS class 1	QoS class 1	QoS class 2
Middle Circle	QoS class 1	QoS class 2	QoS class 3
Outer Circle	QoS class 2	QoS class 2	QoS class 3

Furthermore, if the mobile subscriber that acts each time as client is also a mobile server then the corresponding packets are tagged with the next higher QoS class. By doing so, we favour the exchange of files between mobile servers.

At the AP, the network management module treats the flow-serving queues accordingly. If a queue is full then the dropped packets are discarded and it's up to the end-to-end transport protocol to retransmit the lost packets. Another variation is to use the technique inspired by the Assured Service class [8], as defined by DiffServ, where the extra packets are not dropped but are served by the queue of the lower QoS class.

In fact, the whole procedure is transparent to the mobile subscriber. The AP, according to the position of the mobile subscriber and the serving rate of the respective queue, assigns the appropriate radio resources. Thus, the serving rate at the uplink channel of the mobile server and at the downlink channel from the AP to the mobile clients are the same and equal to the serving rate of the respective queue to which the flow is assigned.

To become more technologically compliant, in UMTS [4] the throughput that a mobile subscriber receives at the uplink is based on the chip rate [5] that he is receiving from the AP. Thus, by differentiating the chip rate for each mobile server we manage to achieve the mechanism proposed in this approach. Note that the generic QoS class, as defined by the UMTS specifications, will be the same for all the traffic flows, e.g. Streaming or Interactive class. Here, we define a lower-level QoS scheme that differentiates further the characteristics of each flow.

We can achieve a similar effect with IEEE WLANs. The upcoming 802.11e [3] standard defines a new algorithm for the MAC layer, in which a WLAN access point that supports 802.11e acts as a point coordinator and polls the stations for data in a manner that is quite similar to the UMTS mechanism. This *Hybrid Coordination Function*, as it's called, ensures certain *Contention Free Periods* in which the access point can issue a *CF-Poll* message to a particular station. This message contains the expected start time and maximum duration that the particular station can have contention-free access to the channel. If the station has data to transmit, it will do so, otherwise the access point will poll another station. Assuming, therefore, 802.11e compliant mobile servers and access points, the access point can allocate to any mobile server the needed wireless resources.

LBS-IS approach:

In the AP, there exist various queues; each serving a flow that corresponds to a file transfer in the context of a specific LBS session. Each queue is assigned to one transmission from a mobile server to another mobile device. Depending on the location of the mobile server and the mobile client, on the type of the mobile client (server or not) and on the network load each moment, a Scheduler module is responsible for assigning different rates at each transmission, based on the first table of the previous section. To do so, the Scheduler module keeps an NxM array, where N is the number of mobile servers and M the mobile subscribers (clients and servers) in the coverage area. In this array, all the above information is kept. The sum of all transmitting rates set by the Scheduler must not exceed the total available bandwidth defined by the radio subsystem (about 6mbps in the WLAN case or up to 2mbps in the UMTS case).

Taking into consideration that we do not cross the maximum cell bandwidth limit, a base station must try to complete as many successful file transactions as possible. We assume that a file is worthless if it does not arrive complete. Assuming that the base station can calculate the average speed of a departing terminal, it has a specific amount of time within which all pending transactions must be finished. If there are many transactions pending, then the "least-significant" ones should be dropped so that the "more-significant" ones can complete. In our algorithm, a mobile server is more important than simple clients. Also, a station that has already received a large percentage of a file has priority with respect to the others (it probably requested earlier in any case).

6. Open Issues – Further Work

So far, we have described a service provision environment where only one AP exists. Our model can easily be extended for more than one AP. In the UMTS case, this can be seen as multiple Base Stations (Node B's) connected to a single RNC. In such an environment and since multiple APs cooperate in order to transfer the files, a different positioning method/technology should be used. Leaving the coverage area of an AP does not mean leaving the service provision area, since the adjacent coverage area of another AP can continue the transmission, the characteristics of which are preserved during the handoff period.

Another issue concerns the incentives given to the mobile subscribers to act as mobile servers. It is true that the mobile subscribers that act as mobile servers will suffer

augmented costs and their devices important power consumption. It is crucial for the operator to provide specific incentives to the subscribers through special pricing schemes so that they actually offer the service. The operator benefits from the provision of such services since his revenues and the network's externalities are increased. A portion of the operator's revenue must return to the mobile subscribers that are also registered as mobile servers, as the actual service providers.

7. Conclusions

In this paper, we have proposed a new business model for service provisioning that can be applied on a wireless environment, based on the evolution of wireless/mobile technologies and the need for a new business model for providing Location-Based Services. The new feature introduced is that the service is provided in a P2P manner, since some mobile subscribers act as mobile servers while other act as mobile clients. An example of such an environment is given, concerning a road traffic monitoring service. The network operator must provide the mobile servers with the appropriate throughput in order to serve the incoming requests. We also take into consideration the relative position of the mobile server and the mobile clients, in order to define the importance and the validity of the content, the need for quick reply and the bounds of the service provision area. Two QoS mechanisms, taken from the Internet paradigm, are used to provide differentiation among the flows in terms of position of server and client and also in terms of client's type (also server or simple client). Differentiated and Integrated Services are introduced and two mechanisms for applying QoS in the wireless environment are proposed.

References

[1] "Top Line Trends In Consumers' Attitudes Towards Location-Based Services In Great Britain, France and Germany," research commissioned by AirFlash Inc. in January 31, 2001, available at: htpp://www.wirelessdevnet.com/channels/lbs/features/airflash.html

[2] C. Ververidis, G. C. Polyzos, "Mobile Marketing Using Location Based Services" in proceedings of the Mobile-Business 2002 Workshop, Athens

[3] Mangold, S. and Choi, S. and May, P. and Klein, O. and Hiertz, G. and Stibor, L. "IEEE 802.11e Wireless LAN for Quality of Service," in Proc. of European Wireless '02, Florence, Italy, February 2002

[4] The 3rd Generation Partnership Project (3GPP), available at: http://www.3gpp.org

[5] H. Holma and A. Toskala, "*WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*," John Wiley, ISBN 0-471-72051-8, September 2000

[6] F. Baumgartner, T. Braun and P. Habegger, "Differentiated Services: A new approach for Quality of Service in the Internet," Proc. of the 8th International Conference on High Performance Networking, Vienna, Austria, 21-25 September 1998, ed. H.R. Van As (Kluwer Academic, Norwell, MA, 1998) pp. 255-273

[7] P.F. Chimento, "Tutorial on QoS support for IP," CTIT Technical report, No. 23 (1998)

[8] J. Heinane, F. Baker, W. Weiss and J. Wroclawski, "Assured forwarding PHB group," RFC 2597 (June 1999)