Abstract—As mobility became the norm rather than the exception, location-based services are playing more of a key role in assisting mobile users. In this paper, we address the challenges of providing location-based services to users in areas of sudden population increases, such as stadiums and traffic jams. A sudden increase in the number of mobile users leads to an increasing demand for the already scarce wireless bandwidth, thus causing dramatic throughput degradation and an increase in connectivity failures. We propose a hybrid model within which a peer-to-peer mode is deployed to assist the cellular-based network whenever a sudden increase in population density is sensed by the base station. Location-based data is migrated to selected elite nodes, thus allowing other peer nodes to get their information locally. This approach is proven, through experimental results, to decrease the query response time and number of query failures.

Index Terms—Location-based services, mobile data management, peer-to-peer networks, service lookup

I. INTRODUCTION

The past years have witnessed an incredible increase in the wireless/mobile user population. The proliferation of mobile devices and the rapid development in wireless communication technology led to the emergence of the pervasive computing paradigm [1]. In such innovative paradigm, mobile users are able to, not only access their data anytime and anywhere, but also benefit from numerous services and context-aware applications from other existing users and devices.

Location-based services are a type of service that emerged as a result of considerable research efforts in the context of location management systems. The location of mobile users is tracked, managed and stored in location servers. Such information is widely used to provide services that are specific to user’s locations. As an example, a user traveling on a city road may be interested in information such as available parking places around its current location, the traffic conditions within one mile ahead, the nearest restaurant to its current place, and others.

Assisting users with location-based services in a wireless environment creates many challenges, among which is the problem of data management. The answer to questions of how to obtain such information within an acceptable delay given the constraints imposed by the wireless environment, attracted many research activities. Constraints include limited bandwidth, limited power resources, low connectivity, and others.

Previous work followed two distinct approaches depending on the wireless network operation mode, namely a wireless cellular approach, and a peer-to-peer approach.

A. The Wireless Cellular Approach

The wireless cellular approach is the most widely used wireless access system. Such approach provides universal coverage, allowing users to move freely from one place to another whilst being able to send and receive data. In such systems, the coverage area is divided into cells, each cell containing a base station responsible for the communication of all mobile users within its area.

Within the cellular based approach, location-based services are usually implemented using the client/server model. Mobile users, clients, submit their queries to remote databases, servers, which reside on the wired network. The communication between the clients and the servers is done through the cell base station resulting in a depletion of battery and wireless resources, increasing monetary costs for cellular connectivity, and high vulnerability to failures of the fixed servers.

Moreover, this approach requires all traffic to and from users to go through the base station, that which may pose a threat of congestion due to the continuous increase in data traffic. This problem is even aggravated in cells experiencing bursting traffic events such as sudden increases in mobile user populations as in stadiums, disaster areas, or traffic jams, thus causing the cell to become congested or a “hot cell”.

A congested cell leads to an increase in query response time, outdated query responses, an increase in the probability of system failures due to limited bandwidth, and dramatic throughput degradation. Many solutions have been proposed to either avoid “hot cell” situations by decreasing the cell size, or by borrowing channels from neighboring less congested cells. The former approach proves itself to be a rather expensive solution, while the latter approach suffers

Sherif G. Aly is with the Department of Computer Science, The American University in Cairo, P.O. Box 2511, 113 Sharia Kasr El Aini, Cairo, Egypt.
Phone: +20 (12) 217-4588. Fax: + (20) (2) 795-7565.
Email: sgamal@aucegypt.edu

Amal Elnahas is with the Department of Computer Science, The German University in Cairo, New Cairo City, Cairo, Egypt.
Phone: +20 (12) 214-7432
Email: amal.elnahas@guc.edu.eg
from inefficiency and higher cost, and does not solve the congestion problem on the server side.

B. The Peer-to-Peer Approach

An alternative to the cellular-based approach is the peer-to-peer approach. An ad-hoc network is a class of peer-to-peer wireless networks. An ad-hoc network is in itself a self-configuring network composed of mobile nodes communicating without the need of any infrastructure. The nodes use short-range network technology, such as Bluetooth or 802.11. Each node acts as a host and a router at the same time, and the union of such nodes forms an arbitrary topology without any centralized administrator, thus relieving the congestion problem.

Ad-hoc networks represent a good alternative in situations where no infrastructure is available like battlefields or disasters rescue areas. They are also significantly less expensive to establish. Moreover, recent studies [2] have shown that in some situations, ad-hoc networks outperform cellular networks in terms of throughput, and delay and power consumption.

The absence of a centralized node in ad-hoc networks led to the inapplicability of the client/server model for data management. Instead, nodes share their information and interchangeably act as information providers and consumers. The problem with such approach is the high failure rate of queries, since not all data may be available in neighboring nodes.

The major drawback of ad-hoc networks lies in the limited transmission range of their nodes, thus being unable to achieve wide-area communications. High vulnerability to frequent route failures, network partitions, and severe throughput degradation for certain traffic conditions also present themselves as ad-hoc network drawbacks.

Although tremendous effort has been conducted to improve both operation modes (cellular and peer-to-peer), none of them is robust enough to replace the other in all situations. Combining the merits of both modes motivated researchers to investigate the idea of using ad-hoc network as an extension to wireless cellular networks [3-8]. Such hybrid approach led to an improvement in overall performance, such as increasing data rate, throughput, enhancing network capacity and load balancing.

In this paper, we propose a hybrid model that can provide mobile users with location-based services in a cellular environment and sustain such services even in situations where a substantial sudden increase in the cell density is detected. Areas of sudden density increases such as stadiums or traffic jams are usually also characterized by low user mobility. The idea we proposed and implemented relies on the conventional client/server model using the cellular mode. Upon detecting an increase in cell population, the base station invocations a peer-to-peer mode by migrating location-based data to selected elite mobile nodes grouped together in what we define as autonomous systems. This allows other querying users to obtain their information locally. This eventually decreases the probability of connectivity failures and increases the number of users that can be accommodated within the cell.

The size of the created ad-hoc network utilizing this peer-to-peer approach is typically smaller than the diameter of a cellular cell, which is typically of the order of one mile. As such, there exists a relatively high probability that the entire ad-hoc network will almost entirely be covered within the area of one single cell. The merits of this idea were investigated through simulation experiments revealing a decrease in number of failed queries and response time.

This article is organized as follows: A review of related work is presented in section II, detection of density spurs is discussed in section III, the arrangement of mobile users and base stations in section IV, the creation of autonomous systems is discussed in section V, service lookup in section VI, the experiments and their results in section VII, and finally, section VIII concludes our work with directions for future work.

II. RELATED WORK

The research in the area of mobile data management was primarily dominated by the classical client/server (or client/proxy/server) model, where mobile users are clients connected to servers residing on a wired network. The challenges faced by researchers were basically due to problems imposed by the wireless environment itself such as variable bandwidth availability, frequent network disconnection, and mobility. Their main objective was to render applications designed for the wired network applicable in the wireless domain. Among the proposed ideas were the use of proxy servers along with data replication and caching techniques for on-demand access [9-14], or data broadcasting for push-based access [15-18].

Apart from the differences in the proposed approaches, they all share the same common feature: users are considered as stand-alone devices with no direct interaction with their neighbors. The full reliance on a fixed node for the communication renders such approaches unscaleable, especially for sudden increase in data traffic. In such situations, high query failure rate, high response time, and degraded throughput can be expected.

On the other side, with the wide-spread adoption of ad-hoc technology, researchers started to move away from the client/server model to the peer-to-peer model. Alternative data management techniques were proposed where users interact and share their data with each others.

A simple opportunistic peer-to-peer approach is introduced in [19] for the dissemination of reports about availability of resources among mobile nodes. The communication is completely based on a peer-to-peer mode and each mobile node propagates reports in exchange for new reports from its peers. This simple approach is not suitable for providing location-based services since it risks high query failure rate due to insufficient information and diversity of users needs.
In [20-21], the MoGATU framework was proposed where users are treated as semi-autonomous peers interacting and sharing data through Bluetooth or ad-hoc 802.11 technologies. Within such framework, data is distributed among a set of nodes, called information providers, and each node has an information manager that maintains information about peers, a data cache, and a user profile. A user querying for a service, sends its query to its information manager which routes it to appropriate information providers. This proposal enables mobile users to share their information in a completely decentralized fashion but does not ensure the availability of all data within the existing users. Moreover, it suffers from the high traffic rate between peer nodes.

In order to be able to provide location-based services to an increasing number of users, we need to combine the merits of both cellular and ad-hoc networks. Integrating both modes to solve the problem of congestion from the routing perspective was first proposed in [6, 7]. Traffic is relayed from congested cells to cells with available bandwidth, thus solving the “hot cell” problem and balancing the load between cells. Deploying the hybrid architecture in data management was introduced in [22, 23], where mobile nodes cooperate to reduce connection cost upon downloading a certain content from the wired network. In [22], each mobile device is assigned a portion of the data, and then shares its portion with other devices through ad-hoc connectivity. In [23], each node took turns, serving as a proxy to download content then share it with its peers. Both approaches are not applicable in location based services since mobile users do not share the same content.

III. DENSITY SPURS

The objective of the research presented herein this article is to sustain lookup services made by mobile users in situations where sudden dense populations occur. The capacity design of geographically dispersed base stations primarily depends on handling normal loads within the indicated area.

During some scenarios however, sudden density spurs can occur within the geographical areas, eventually incapacitating base stations from being able to provide services at the desired or even sub-desired quality of service levels. The probability of occurrence of sudden density spurs varies by nature from one geographic area to another, and from one scenario to another.

A. Anticipated Spurs

As an example, stadiums usually experience infrequent yet unexpected sudden density increases within that geographic location. A liberal approach for base station design would be to design the capacity of the base stations for a worst case scenario, where the stadium is at maximum capacity. However, this is intuitively at the expense of significant cost. On the other hand, the cost incurred for a worst case design scenario as such is even more unjustified when the capacity of the users within the stadium falls to minimal levels at times of inactivity. Such example demonstrates that although sudden density spurs are expected, and the maximum capacity is known, a significant unjustified cost incurred in the design of the base station must be somehow reduced.

B. Sporadic Spurs

Numerous examples exist for situations where sporadic and unanticipated increases in density can happen. A primary example includes traffic jams. Base station capacity along highways or city roads is usually designed to handle normal load, including the load made by the surrounding residential and commercial blocks, as well as the load made by moving traffic.

However, this is an example where designing base stations for maximum capacity to provide the desired quality of service in all scenarios is almost impossible. Traffic jams are an example of highly sporadic events. At some point in time, jams can occur at a given city block or highway segment, and at other times, they can happen elsewhere. Although the occurrence of the traffic jam itself may be frequent and inevitable, its whereabouts is highly dependent on usually unanticipated events.

C. Density Spur Detection

The calculation of density in itself varies from one context to another; where density calculation is rather a function of both the number of mobile users within the coverage range of a base station of a given maximum capacity.

The detection of sudden density increase is rather simple and depends on sampling across time. It depends on calculating the current density increase within the base station and comparing it to historical densities in the short term. As indicated in Figure 1, we can detect a density spur if the density change \( i_k - i_{k-1} \) increased by steps of predefined values at the checking point \( t_k \). Such step increases, as well as the checking interval could be customized from one base station to another. In scenarios where base station coverage overlaps, density calculations are only made upon those users who are registered with one of the base stations.

![Figure 1 Density Increase](image)

The density spur detection mechanism employed within this section can be easily enhanced and modified without affection the remaining design of our system. A more
elaborate density spur detection algorithm may certainly be used.

One approach of handling large numbers of users requesting services within base stations is to design base stations for worst case scenarios, another intuitive approach is to offload base stations themselves from communication traffic at times of sudden density increases, at the expense of mobile user resources residing within the geographic area of the base station. We have utilized the latter approach within this research as a means of sustaining service lookup with a desired quality of service.

IV. MOBILE USERS AND BASE STATIONS

A. Base Stations

Each base station is responsible for keeping track of mobile users within its geographical service area. A mobile user entering the geographical area of a base station will register itself with this base station. Registration includes not only the user’s identification, but also its resources as indicated in the following section.

All traffic made by registered mobile users goes through the base station in which the mobile users are registered. Base station traffic may include both voice, as well as service lookup such as requests to locate the nearest hospital, or the closest airport.

However, when sudden density spurs occur, the base station’s ability to serve such significant number of mobile users quickly diminishes. The processing, storage, and communication resources of the base station are potentially liable to hasty consumption by various mobile users attempting to initiate both voice and service lookup.

One rudimentary solution to overloading is to simply deny mobile user requests that exceed the ability of the base station to provide quality of service. However, this approach simply guarantees a quality of service for a small portion of the users that had the privilege of being serviced by the base station, but as to the other users attempting to initiate any form of traffic, they receive no service at all.

Within this research, we focus on service lookup, and allow mobile users to successfully lookup requests bound by some form of quality of service guarantee. This is achieved even if the base station itself will usually in a similar scenario deny such requests from going through, or simply provides the service at an inferior service level.

B. Mobile Users

Mobile users roaming around a network naturally enter and leave various base station coverage areas as they roam around. We would like to utilize a subset of the dense multitudes of resource-limited mobile users within the base station coverage area. Such utilization is made by forming clusters of resources capable of caching and processing lookup data, and eventually supporting service lookup without the need to go through the base station itself. As such, the base station is relieved from potential processing and communication overload incurred by the dense existence of mobile users within its service area.

We are primarily interested in four types of resources provided by various mobile users, namely: processing, primary storage capacity, secondary storage capacity, and communication bandwidth. Each mobile user registering with a base station is quantitatively rated according to the four types of resources of interest indicated above. We can quantify the resources of mobile users in the form of a single number representing the resources of such user. Both primary and secondary storage are quantified in terms of bytes. Communication bandwidth is quantified in terms of bits per second. As relates to processing power, we are not interested in a very accurate quantification of processing powers; however, a heuristic-based approach is followed in estimating processing power. Processing power is assumed to be directly related to primary, secondary, and communication abilities of a mobile user. Eventually, a single number formed by summing up each of the quantifications of primary and secondary storage, communications bandwidth, and processing power. Such resource quantification is part of a mobile user’s registration at the base station. Registering such mobile user resources at the base station presents negligible overhead.

A selected elite of the mobile users registered with the base station will be designated as super nodes, and will be utilized in the formation of clusters, what we call autonomous systems, capable of locally responding to service lookup.

C. Super Nodes

Super nodes are an elite subset of the total set of mobile users registered within a given base station as shown in Figure 2. It is the super node’s resources that will be utilized in order to provide local service lookups to other mobile users, including other super nodes registered at the base station also.

Figure 2 Super Nodes as a Subset of Mobile Users

A mobile host manager residing at the base station is responsible for the selection of super nodes. As previously indicated, not only will a mobile user register its identification with a base station, but it will also register its resources. The selection of super nodes will be made
according to a quantification of such resources.

**D. Mobile Host Manager**

Once a density spur is detected by the base station, the mobile host manager will immediately attempt to identify the super nodes from amongst the already registered mobile users. In our case, we select an elite twenty percent of the already registered mobile users as super nodes. However, although a mobile user may qualify as one of the indicated twenty percent, it may deny its participation as being a super node. The percent of mobile users to be selected is configurable from one base station to another.

It is worth observing that, as the number of super nodes selected from amongst the overall set of mobile users within a given coverage area increases, serious drawbacks start to exist. First of all, the amount of traffic incurred within the coverage area of the base station and amongst super nodes will significantly increase as the number of super nodes increases. Second, not only does this mean extra utilization of the bandwidth of each super node, but it also has serious implications on the utilization of very valuable power resources of already resource limited devices. Third, the overhead of maintaining data in the super nodes also increases. To this extent, the number of super nodes should not exceed beyond an approximate twenty percent of overall users.

**E. Dispersion**

Selecting super nodes as a percentage of mobile users with the highest most resources may not be a good choice after all. There exists a direct relationship between the quantification of a mobile user’s resources and its cost price. In a scenario like the stadium scenario, users with expensive and resourceful devices, primarily super node candidates, may very well be concentrated into one single area of the stadium, the area with more expensive tickets.

The concentration of all super nodes, those nodes that will service other mobile users, into one small geographical area of the base station’s coverage area significantly impacts communication latency between mobile users and super nodes. As a best case scenario, we would like super nodes to be equally dispersed inside the coverage area of the base station as shown in Figure 3, and at the same time we need to maintain the super nodes as resourceful devices capable of servicing others. It is even more difficult to guarantee this dispersion due to the mobility of users.

![Coverage Area](image)

**Figure 3 Good and Bad Distributions of Super Nodes**

In order to guarantee dispersion of super nodes to some extent, while maintaining the selection to be amongst the most resourceful devices, we select the indicated twenty percent of mobile users from the highest ten percent, the following ten percent, and the third ten percent of mobile users according to the following table:

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Percent Selection from Overall Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest Ten Percent</td>
<td>10</td>
</tr>
<tr>
<td>Second Highest Ten Percent</td>
<td>7</td>
</tr>
<tr>
<td>Third Highest Ten Percent</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
</tr>
</tbody>
</table>

Eventually, we select the indicated total of twenty percent of the overall mobile users, but we do not perform the selection entirely from the most resourceful devices alone. As shown in the Table I, the twenty percent designated as super nodes are composed of the most resourceful ten percent of users, followed by the most resourceful seven percent of the second ten percent, followed by the most resourceful three percent of users from the third ten percent, thus accumulating to a total of twenty percent of the entire mobile users. This utilized heuristic supports the dispersion of super nodes within the congested area.

The percent selected from the various rankings above is customizable and could be modified from one base station to another. However, the lower the ranking is, the less the resources of the mobile user, and the less selection should be made from amongst that specific rank.

The percent selection we chose above was based on selecting the entire highest ten percent, and the remaining ten percent was selected from the second and third rankings by an approximate 2:1 ratio.

The overhead associated with the selection of super nodes is also a negligible overhead that involves simple sorting and selection of high resource nodes.

**V. AUTONOMOUS SYSTEMS**

Upon detection of a density spur, and as previously indicated, the base station, with the help of a mobile host manager, will perform the selection of super nodes from amongst the set of mobile users already registered within the base station. Super nodes are clustered together into identical entities called autonomous systems (AS). Each autonomous system will cache on a best effort basis lookup data from an external database usually queried by the base station to perform lookup requests.

**A. Autonomous System Formation**

Each autonomous system is composed of a maximum of a predetermined number of super nodes. The maximum
number of super nodes making up an autonomous system is a configuration parameter in the base station. However, the last autonomous system to be formed may not find enough super nodes to form a fully occupied one.

As a heuristic, if the remaining number of super nodes is incapable of forming an autonomous system with at least half of the maximum allowed size, the entire autonomous system is discarded. This way, we only rely on autonomous systems operating at maximum capacity, except in the case of the last one, which must be operating at least with half the maximum allowable capacity as defined in the base station. The formation of autonomous systems is illustrated in Figure 4.

![Figure 4 Super Node Formations into Autonomous Systems (AS)](image)

Similar to their formation, autonomous systems must be updated whenever super nodes un-register with the base station. Whenever a super node leaves the base station’s geographical coverage area, it can no longer be utilized as a super node within that base station. Before un-registration, the data of the super node is acquired, and another mobile user is selected by the base station, and replaces the departing super node.

A less aggressive mechanism in autonomous system updates allows autonomous systems to be entirely replaced whenever the number of super nodes composing the autonomous system falls beyond a certain threshold. Such technique allows the base station to recover from the loss of super nodes by bulk rather than one at a time.

B. Autonomous System Population

After successfully forming the autonomous systems from the set of selected super nodes, the base station initiates a one time data transfer of lookup data into each autonomous system. This way, the lookup data is cached on a best effort basis into each autonomous system residing within the geographic coverage of the base station itself.

The base station queries data from the service database in chunks, and attempts to populate each autonomous system to its maximum storage capacity as shown in Figure 5. The amount of data stored may differ from one autonomous system to another as the capacity of each autonomous system is dependent on a couple of factors, namely:

1. The number of super nodes in each autonomous system.
2. The amount of storage resources available in each autonomous system. This varies according to the resources of the super nodes forming the autonomous system.
3. The size of each data record itself.

![Figure 5 Autonomous Systems Population](image)

Given the following:

- $n$ is the number of super nodes in an autonomous system,
- $s_i$ is the available shared storage capacity of a super node $i$ in the autonomous system,
- $m$ is the total number of records in the database,
- $r_j$ is the record size in bytes, and $x$ bar is the average record size in the database,
- The average number of database records in an autonomous system, $NR$, is governed by the following equation:

$$NR = \left\lfloor \frac{\sum_{i=0}^{n-1} S_i}{x} \right\rfloor$$

Where

$$x = \frac{\sum_{j=0}^{m-1} r_j}{m}$$

C. Record Selection

The question that primarily raises itself is which database records to migrate to the various autonomous systems. In a worst case scenario, most of the database records populated inside the autonomous systems could be data that is rarely queried for in the first place. As a best case scenario, the autonomous systems will contain the most popular data being queried for.

As such, there are three primary techniques for selecting database records to populate into the autonomous systems:
1. **Statistical Selection:** Given the database itself contains statistics of most popular requested records; such records can be the primary choice of population into the autonomous systems. However, this technique is highly dependent on the presence of statistics in the database itself, and that the statistics are meaningful in the context we are operating within.

2. **Sequential Selection:** Records from the database could be selected starting from the first record until the maximum capacity of the autonomous system is encountered. This technique however is simple, yet has a serious drawback in the sense that the selected records could be highly unpopular in terms of choice.

3. **Distribution Based Selection:** Records from the database could be selected according to a given mathematical distribution. In the simple sense, data could be selected at random to be populated within the autonomous systems. This is the most flexible technique, assumes a uniform access to records, and that no statistical information should be present in the database. Furthermore, different autonomous systems will have different sets of data from the database, thus serving different sets of querying mobile users. This is the recommended technique of choice.

Although the download of records to autonomous systems presents itself as the most impressing overhead so far, the objective of creating this peer-to-peer mechanism should not be overlooked. The primary objective of this mechanism is for it to be deployed as a base station life-saving mechanism at times of sudden dense populations. Such download overhead could be easily justified. Furthermore, the downloading of records to autonomous systems need not occur all at once, but rather can occur over an extended duration of time.

**D. Data Indexing**

In order to facilitate subsequent data lookups within an autonomous system, the base station will add the data into the autonomous systems in sorted order of the lookup field. Each super node within the autonomous system will contain an index indicating the lookup field of the first record inside the super node, and the lookup field of the last record.

Using this indexing information, the sorting of data inside the autonomous systems, and the logical linking of autonomous systems in a linked list topology, the lookup of data can be significantly simplified as indicated later in the service lookup section.

VI. **SERVICE LOOKUP**

Any mobile user within the geographical area of the base station can request multitudes of lookup services along with the usual phone conversations. Lookup services include locating the closest restaurants, the closest hospital with a given specialty, the arrival times of airplanes or trains, and so on.

Service lookup can even be sensitive to events taking place, including historical scores of currently playing teams, or scores of similar teams playing during the same time, or during traffic jams, finding alternative routes to the current route. In this section, we will demonstrate how a lookup query locating the nearest location of an airport or hospital or school is achieved using our mechanism.

Under normal circumstances, every lookup request initiated by a mobile user will usually go through the base station, to the lookup database, back to the base station, and then to the mobile user. However, under heavy load scenarios, such as when the density within a given base station’s geographical coverage suddenly increases, substantial load is incurred by the base station in an attempt to service such lookup requests.

Besides processing delays, each base station has a limited bandwidth utilized in both the transmission and reception of query data. Under load scenarios such as the one indicated, substantial queueing delays start to appear, thus affecting the quality of service provided to the query requests.

**A. Autonomous System Assignment**

However, utilizing our mechanism, lookup data is either partially or fully cached within the geographical coverage of the base station itself. Such data is cached within the resources of the selected elite super nodes. Once a mobile user initiates a service lookup, and after detection of a density spur, the base station checks to see if the mobile user already has an assigned autonomous system. If no autonomous system is assigned to the querying mobile user, the base station will assign the user an autonomous system.

Autonomous systems are assigned in a circular fashion; autonomous systems are assigned to mobile users one after the other sequentially, until the last autonomous system is reached, the first autonomous system is assigned one more time to another user. This way, we can promote fairness in the assignment of autonomous systems to mobile users, and thus promote fair load amongst various autonomous systems.

**B. Service Querying**

Once a mobile user is assigned an autonomous system, the assigned autonomous system of the mobile user itself is queried for the data to be looked up. Each super node in the user’s assigned autonomous system is checked one after the other until the required data is to be found. As previously indicated, each super node has an index of the first and last data records stored within it to avoid unnecessary search and overloading of super node resources.

If the data is found within the mobile user’s assigned autonomous system, the result is immediately returned by the super node within which the result was found to the querying
mobile user. In such scenario, when the data is found locally in an autonomous system, the base station itself is relieved from any processing and communication overhead to perform the lookup query.

On the other hand, if the result of the query was not found within the autonomous system of the querying mobile user, the query will then have to normally go through the base station itself and the result returned back to the querying user. This means that the more data we can locally store within the autonomous systems, the more relief is made to the base station itself.

There are many techniques by which autonomous systems can be assigned to various users. The assignment of autonomous systems to mobile users within the densely populated area, along with the querying process is indicated in the flow chart of Figure 6. It is worthy to note how we check for the existence of service lookup data within the autonomous systems first in an attempt to offload base stations. The base station’s services are only needed when the service lookup data is not found within the autonomous systems.

C. Routing

Many forms of communication need to take place within this system:

1. Mobile users need to issue query requests to their autonomous system.
2. When query results are found, super nodes need to be able to return back query results to mobile users.
3. When query results are not found at a given super node of an autonomous system, the super node where the query was not found needs to communicate with the next super node of the autonomous system to relay the query request.

Communication is either achieved through a direct wireless communication between various mobile users or super nodes, or when the transmission range is not sufficient, mobile users will relay messages using one of the well known ad-hoc routing algorithms, like Dynamic Source Routing protocol (DSR) that efficiently uses the available bandwidth, eventually reaching the destination. A minimum overhead is incurred because of the low user mobility rates.

D. Data Privacy

Data distributed amongst various autonomous systems is most certainly an issue of security concern. However, the data we are discussing here, relevant to locating services such as the nearest hospital or police station, is by all means public data accessible to all mobile users. Data requiring higher security standards cannot be used within this mechanism in its current form. Such security-demanding data must be used within here along with security measures including data and communication encryption.

E. Autonomous System update

Autonomous system composition is subject to change. Super nodes may easily leave the coverage area of the base station, and hence deregister. This action will cause the selection of another super node from amongst registered mobile users to replace the departing super node.

Each departing super node has an index of data records it previously stored. This index will help in the determination of the exact database records to download to the replacement super node.

Another approach towards autonomous system update can rely on lazy replacement of super nodes. A single departing super node may not necessarily prompt immediate replacement. Rather, only after a certain threshold of departing super nodes is achieved will the base station attempt to restructure the autonomous systems one more time.

VII. THE EXPERIMENTS

To further prove our concept, we conducted several experiments to measure the behavior of the mechanism under various conditions.
A. Hit Percentage

Hit percentage is defined as the percentage of queries that were successfully found locally within the autonomous systems out of the overall number of queries performed by mobile users. In this experiment, we utilized a lookup database of a fixed size; ten thousand lookup queries were generated according to an exponential distribution. We varied the size of the autonomous systems starting from zero super nodes up until one hundred super nodes in incremental steps of five. The objective was to monitor the behavior of the percent hits as the size of the autonomous systems increased. The following result was obtained:

![Figure 7 Hit Percentage]

From the graph in Figure 7 above, we see a semi-linear relationship between the autonomous system size and the hit percentage. Clearly, when the autonomous system size is zero, all queries go through the base station, and we observe a zero hit percentage. As the number of super nodes in the autonomous systems increases, the hit percentage also increases almost linearly. A saturation point in the hit percentage is reached when the total amount of storage resources of the autonomous systems is capable of caching the entire lookup database. At that point in time, all queries are performed locally, and the hit percentage is one hundred percent.

The total amount of storage resources of the autonomous systems is a function of both the autonomous system size, and the amount of storage resources of the super nodes forming the autonomous system itself. Of course, a best case scenario would be to entirely cache the lookup database within the autonomous systems, however, as previously indicated, this is at the expense of the super nodes, their resources, and internal traffic within the coverage area of the base station itself.

B. Query Latency without Density Sensitivity

The second experiment we conducted was intended to measure the query latency in a situation where we have a density spur, but however, without utilizing our mechanism. We define query latency as the amount of time taken since the initiation of a query request, until the result is returned back to the querying user.

In this experiment, ten thousand mobile users arriving at an exponential distribution were utilized. Each mobile user generated a query, and the query latency was measured. Obviously, each generated query went through their base station of a predefined processing and bandwidth capability. Our objective was to measure the query latency without applying the density sensitivity mechanism explained in this article. The following result was achieved:

![Figure 8 Query Latency without Density Sensitivity]

From the curve in Figure 8 above, it becomes apparent that initially, the base station is capable of servicing the querying mobile users without any substantial latency increase. As long as the base station is capable of servicing the mobile users requesting queries, the query latency remains around one hundred milliseconds as indicated from the values of the logarithmic y axis of the chart above.

As more mobile users perform queries, queuing delays start to appear at the base station. The base station becomes incapable of servicing the querying mobile users in due time. The query latency reaches values close to ten thousand milliseconds or more, approximately one hundred times above the normal value of query latency indicated at the beginning of the curve. Such latency delay can of course be handled in two primary techniques as earlier indicated, either to block query requests beyond the capacity of the base station, or to find more innovative techniques to handle such load as indicated in this article.

C. Query Latency with Density Sensitivity

The third experiment we conducted was intended to measure the query latency in a situation where we have a density spur, and with our mechanism in place. The same definition of query latency is utilized. The results of this experiment actually prove that we were able to accommodate for more mobile users within a densely populated area. Within this experiment, the same number of mobile users is used, except that now, autonomous systems service those mobile users at first, thus offloading the base station itself.

Ten thousand mobile users also arriving at an exponential distribution were utilized. We utilized autonomous systems of a maximum size of fifty super nodes. The same database size was utilized as in the first and second experiments. Each
mobile user generated a query, and the query latency was measured. If the query made by a given mobile user was found in its assigned autonomous system, the data is looked up from the autonomous system and the result is returned back to the querying mobile user. Otherwise, the query request had to go through the base station. Our objective was to measure the query latency with our mechanism in place. The following result was achieved.

From the curve in Figure 9 above, it becomes apparent how a significant portion of the queries are serviced by the autonomous systems, without the need to go through the base station. The query latency is significantly lower than the latency associated with a query going through the base station, even under light load conditions. For the rest of the queries that went through the base station, the query latency is approximately one hundred milliseconds. The important thing to note is that the base station does not experience significant querying delays as shown in Figure 8.

The results of this experiment demonstrate two primary gains: The first being a significant reduction in query latency delay by servicing the queries locally via the autonomous systems, without the need to go through the base station. The query latency is significantly lower than the latency associated with a query going through the base station, even under light load conditions. For the rest of the queries that went through the base station, the query latency is approximately one hundred milliseconds. The important thing to note is that the base station does not experience significant querying delays as shown in Figure 8.

The results of this experiment demonstrate two primary gains: The first being a significant reduction in query latency delay by servicing the queries locally via the autonomous systems, without the need to go through the base station. The query latency is significantly lower than the latency associated with a query going through the base station, even under light load conditions. For the rest of the queries that went through the base station, the query latency is approximately one hundred milliseconds. The important thing to note is that the base station does not experience significant querying delays as shown in Figure 8.

The following graph in Figure 10 demonstrates the reduction in latency achieved by applying our mechanism. The data points in the chart are achieved as the difference between the points when the mechanism was not applied, and when the mechanism was applied as explained in the previous two experiments. It is apparent in the figure how the reduction on query latency is achieved most when base station experienced query delays due to the large volume of service lookups being attempted.

Figure 9 Query Latency with Density Sensitivity

From the curve in Figure 9 above, it becomes apparent how a significant portion of the queries are serviced by the autonomous systems, without the need to go through the base station. The query latency is significantly lower than the latency associated with a query going through the base station, even under light load conditions. For the rest of the queries that went through the base station, the query latency is approximately one hundred milliseconds. The important thing to note is that the base station does not experience significant querying delays as shown in Figure 8.

The results of this experiment demonstrate two primary gains: The first being a significant reduction in query latency delay by servicing the queries locally via the autonomous systems, without the need to go through the base station. The query latency is significantly lower than the latency associated with a query going through the base station, even under light load conditions. For the rest of the queries that went through the base station, the query latency is approximately one hundred milliseconds. The important thing to note is that the base station does not experience significant querying delays as shown in Figure 8.

The following graph in Figure 10 demonstrates the reduction in latency achieved by applying our mechanism. The data points in the chart are achieved as the difference between the points when the mechanism was not applied, and when the mechanism was applied as explained in the previous two experiments. It is apparent in the figure how the reduction on query latency is achieved most when base station experienced query delays due to the large volume of service lookups being attempted.

Figure 10 Latency Reduction

D. Autonomous Systems Tolerance to Increasing Service Lookup

The curve in Figure 8 above does not indicate by any means the same latency behavior if the number of mobile users were to increase beyond the ability of the autonomous systems to gracefully respond to service lookup. If we were to increase the number of mobile users so dramatically in this experiment one more time, queuing delay within our peer-to-peer mechanism, similar to the ones in Figure 8 will most certainly also appear as a normal reaction to congestion. Such congestion within autonomous systems however must be controlled to preserve the performance of the mobile users within the autonomous system itself. Although the possibility of such scenario happening is valid, yet, the probability of its occurrence tends to be low due to the natural limitation on the amount of mobile users that can exist within a given physical boundary. This is subject to future research.

VIII. CONCLUSION

It is possible to substantially decrease query latency and failures associated with service lookups in cells experiencing a sudden increase in subscriber density through the migration and caching of service lookup data within an ad-hoc structure of elite mobile nodes called autonomous systems.

Initially, each mobile user registering with a base station can optionally register its available resources with the base station itself. Upon detection of a sudden increase in mobile user density, base stations initiate the formation of autonomous systems composed of a selection of elite mobile users called super nodes. Service lookup data is then migrated to autonomous systems on a best effort basis. There are multiple techniques for the selection of records to migrate into autonomous systems, from those, we select and justify the usage of a distribution based selection strategy.

Under high density conditions, service lookup queries generated by mobile users are initially handed over to an assigned autonomous system. Finding the required service lookup data within the autonomous system implies a query hit and implies that the base station was relieved from handling such query.
We performed multiple experiments to demonstrate the hit percentage as the size of autonomous systems increase. Experiments also included the behavior of query latency under dense cell conditions with and without the utilization of our density-sensitive mechanism. The utilization of our mechanism substantially decreased query latency, even with the presence of a high density within the cell itself.

Utilizing the mechanism stated herein this article, the number of users existing within a cell performing service lookup functions can substantially increase without query latency degradation, or even service interruptions.

We are currently investigating the utilization of user profiles to predict anticipated service lookups. We are also investigating prediction of service data migration needs.

REFERENCES


Sherif G. Aly received his B.S. degree in Computer Science from the American University in Cairo, Egypt, in 1996. He then received his M.S. and Doctor of Science degrees in Computer Science from the George Washington University in 1998 and 2000 respectively. He worked for IBM during 1996, and later taught at the George Washington University from 1997 to 2000 where he was nominated for the Trachtenberg prize-teaching award for his current scholarship and scholarly debate. He spent two years as a guest researcher for the National Institute of Standards and Technology at Gaithersburg, Maryland from 1998 to 2000. Dr. Aly also worked as a research scientist at Telcordia Technologies in Morristown, New Jersey, in the field of Internet Service Management Research, and as a Senior Member of Technical Staff at General Dynamics Network Systems. He also consulted for Mentor Graphics and taught at the German University in Cairo. He is currently a faculty member at the Department of Computer Science at the American University in Cairo. Dr. Aly published numerous papers in the area of distributed systems, multimedia, digital design, and programming languages. His current research interests include pervasive systems, programming languages, multimedia, directory enabled networks, and image processing. Dr. Aly is a member of IEEE.

Amal A. Elnahas received her B.Sc with honors and M.Sc. degrees in Computer Engineering from the Alexandria University, Egypt, in 1986 and 1990 respectively. She received her Ph. D. degree as a joint supervision between Alexandria University and The University of Texas at Austin in 1996. Dr. Elnahas worked as a teaching and research assistant in Computer Engineering since 1986 in Alexandria University. In 1995, she was a visiting scholar at The University of Texas at Austin for two years. She then worked as an Assistant Professor in the department of Computer Engineering in Alexandria University from 1996 till 2000. She spent three sabbatical years working in Ain Shams University, Cairo, Egypt as Assistant Professor. In 2003, she joined the German University in Cairo as Associate Professor in the department of Computer Science and Engineering. Dr. Elnahas has a various number of journal and conference papers in the field of network protocols, mobile computing, moving-object databases, and ad-hoc networks. Her current research interests include mobile and wireless network protocols with an emphasis on quality of service guarantees and location management, ad-hoc networks, location-based services, and heterogeneous networks.