Computer Communications 33 (2010) 92–102

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Achieving optimal data storage position in wireless sensor networks

Zhaochun Yu [a,b,c], Bin Xiao [b,*], Shuigeng Zhou [a]

[a] Department of Computer Science and Engineering, Fudan University, China
[b] Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[c] Data Center (Shanghai), Industrial and Commercial Bank of China, China

ARTICLE INFO

Article history:
Received 27 April 2008
Received in revised form 20 May 2009
Accepted 5 August 2009
Available online 15 August 2009

Keywords:
Wireless sensor networks
Optimal data storage
Data rate
Geographical location

ABSTRACT

Data storage in wireless sensor networks (WSNs) involves *producers* (such as sensor nodes) storing in storage positions a large amount of data which they have collected and *consumers* (e.g., base stations, users, and sensor nodes) then retrieving that data. When addressing this issue, previous work failed to utilize data rates and locations of multiple producers and consumers to determine optimal data storage positions to be communication cost-effective in a mesh network topology. In this paper, we first formalize the data storage problem into a *one*-to-*one* (*one* producer and *one* consumer) model and a *many*-to-*many* (*m* producers and *n* consumers) model with the goal of minimizing the total energy cost. Based on above models, we propose optimal data storage (ODS) algorithms that can produce global optimal data storage position in linear, grid, and mesh network topologies. To reduce the computation of ODS in the mesh network topology, we present a near-optimal data storage (NDS) algorithm, which is an approximation algorithm and can obtain a local optimal position. Both ODS and NDS are locality-aware and are able to adjust the storage position adaptively to minimize energy consumption. Simulation results show that NDS not only provides substantial cost benefit over centralized data storage (CDS) and geographic hash table (GHT), but performs as well as ODS in over 75% cases.

1. Introduction

Data storage in wireless sensor networks (WSNs) [3,14] involves producers (such as sensor nodes) storing in storage positions a large amount of data which they have collected and consumers (e.g., base stations, users, and sensor nodes) then retrieving that data. The data storage strategy, including the decision as to where to place storage positions for producers and the provision of responses to consumers' queries, is critical in WSNs. A poorly designed data storage strategy will increase communication overheads, dissipate valuable energy and reduce the lifespan of battery-powered sensor networks. In contrast, a good storage strategy can tremendously reduce the energy consumption for data storage and retrieval, minimize query processing delays and prolong the lifespan of a sensor network. Further, more desirable are strategies that can place data adaptively so as to minimize the costs of storage and query.

Two main factors that impact data storage-related communication cost are the data rates of producers and consumers and their path distances to the storage node. The data rate of producers denotes the data producing rate from producers. The data rate of consumers denotes the data querying rate from consumers. The data rate usually does not change in a fixed application-specific time interval. For example, where there is only one producer and one consumer, data would be stored closer to the consumer rather than the producer when the query rate is higher than the data producing rate, and vise versa. In a real sensor network, the closer the storage node is to the producers and consumers, that is, the shorter the hop distance, the cheaper it is to store and query a fixed quantity of data. An effective way to do this is to place data adaptively according to network state (e.g., locations of nodes requesting data and their data rates) so that the communication cost is minimal once the data storage position is placed.

Although these two factors had been investigated, previous work mainly addressed the storage problem by treating the sensor network as a tree structure, in which the base station [10,15,19] is normally treated as the storage node or the only consumer. In the tree structure, the data rates of producers and the query rate of the base station are known or at least predictable and communication cost can be optimized as data storage placement is simply a response to data volumes. In a mesh network topology, however, there are potentially multiple producers and consumers all seeking to exploit one event simultaneously. In this scenario, some work has focused on the geographical locations [18] of producers and consumers but given no attention to the issue of the data rates [5,8,13]. Once the network topology is fixed, the storage node

* Corresponding author. Tel.: +852 27667270.
E-mail addresses: zcyu@fudan.edu.cn (Z. Yu), csbxiao@comp.polyu.edu.hk (B. Xiao), sgzhou@fudan.edu.cn (S. Zhou).

0140-3664/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2009.08.005

position is fixed too, which does not apply adaptive storage principle to reduce energy consumption in sensor networks. The drawback is that the storage node positions may not be location-aware nor adaptive to network state.

Rather than determining n storage locations[1] for an event, we focus on the single-node storage problem in a wireless sensor network with a mesh topology, where information collected from all producers is sent to a storage node and all consumers retrieve information from there. Similar to previous centralized data storage approaches like [10,13,19], the single location can be event-oriented, i.e., one storage node to store data related to a particular event. Data load is generally asymmetric in a mesh network topology, e.g., for the security purpose [20]. Such an unbalanced data volume causes an uneven energy consumption distribution among different sensor nodes. Since traffic load and energy consumption of each node are location-dependent and rate-dependent, the network lifetime can be shortened by nodes with greater energy consumption. Therefore, storage node placement scheme can have considerable impact on the network lifetime.

In this paper, we propose an optimal data storage (ODS) strategy in a wireless sensor network that allows the storage location to adaptively change, in response to both the geographical locations of producers and consumers and the data rates at which data are being exchanged. This strategy minimizes the energy consumption of the total network, and decreases the delay for message exchange between producers (or consumers) and the storage node. That is, storage position varies adaptively in response to data rates of nodes and their geographical locations. Specifically, we design and implement our adaptive data storage strategy to determine the total communication cost in the *one*-to-*one* (*one* producer and *one* consumer) and *many*-to-*many* (m producers and n consumers) models.

In the *one*-to-*one* model, ODS algorithm can generate an optimal storage location with minimum global communication cost. In the *many*-to-*many* model, multiple producer and consumer nodes can distribute in linear, grid, or mesh network topologies. Given each distribution, we propose distinct ODS algorithms to locate the globally optimum storage position. We show that to determine the optimal storage position, the ODS algorithm complexity is $O(1)$ in the *one*-to-*one* model, $O(m+n)$ in the linear and grid topologies for m producers and n consumers. In other words, to get an optimal data storage location for sensors regularly distributed, ODS only requires *constant* computation time when $m+n$ is much smaller than N, the total number of nodes in a sensor network. In the mesh network topology, ODS becomes a greedy algorithm and its time complexity is $O(N*(m+n))$. To reduce its computational overheads, we aim to get an approximation solution from the geometry theory and propose NDS, a near-optimal data storage strategy, an alternative to ODS. NDS reduces the computation complexity to be a fraction $\pi R^2/S$ of ODS where R is the sensor transmission range and S is the sensor distribution area.

We conduct extensive simulations to show that, compared to centralized data storage (CDS) [10,19] and geographic hash table (GHT) [13], ODS can greatly reduce the energy consumption as well as minimize delays in the data exchange process. NDS is also very efficient to generate the optimal storage position as ODS in more than 75% simulation cases.

The rest of this paper is organized as follows. Section 2 discusses related work. In Section 3, we describe terminology and assumptions to formulate the data storage problem to save the communication energy cost. In Section 4, we implement ODS in the *one*-to-*one* model and show its performance evaluation. Section 5 presents the ODS and NDS algorithms in the *many*-to-*many* model and Section 6 illustrates simulation results. Section 7 offers our conclusion.

[1] As denoted in [12], it is an NP-hard problem for the file multiple copies allocation in computer networks. Thus, in this paper, we limit our discussion for optimal but efficient storage location in a single node in wireless sensor networks.

2. Related work

Data storage in wireless sensor networks are addressed in either tree structure or mesh network topologies. Tree structures feature only one consumer (usually the base station at the root) and multiple producers and do not take geographical location information into account when determining data storage placement. In contrast, a mesh network involves multiple producers and consumers yet the dominant approach in the design of data storage to minimize communication overhead has been given to emphasize geographical locations with little attention to the data rates.

The tree structure offers a range of data storage strategies. Data can be stored and processed in a centralized server or in a distributed manner. In centralized data storage (CDS) approaches such as COUGAR [19] and TinyDB [10], all sensor nodes feed their data to the base station at the root of the tree. Scoop [4] does collect statistics about data, queries, and network conditions and uses them to dynamically change an in-network storage policy. However, it collects these statistics only periodically and uses a greedy algorithm to compute the optimal storage position. Such a try-and-test greedy algorithm is tremendously complex, making it infeasible for a large-scale sensor network. Sheng et al. [15] utilized data rates, query rates, and compression ratio to determine storage placement and introduced storage nodes to alleviate the heavy load of transmitting all data to a central place (the base station). They proposed the optimal placement of multiple storage nodes but it can be only applied in a tree topology. The local data storage (LDS) in the tree structure requires that sensor nodes store data locally in their own memory. Queries are flooded to all nodes in the network (or at least to all nodes that could possibly hold the relevant data) and nodes holding the appropriate data then reply. Apart from flooding, queries can also be disseminated by direct diffusion [7], in which query costs are reduced by using in-network aggregation.

In the mesh network topology, the typically proposed brokerage rules are data centric storage (DCS) [16] and geographic hash table (GHT) [13]. Both methods combine the idea of a distributed hash table (DHT) in peer-to-peer system with that of geographic naming and routing. They use geographical locations as reservoirs where data are hashed to and retrieved from. DIM [8] presents a distributed index for multi-dimensional data that uses special locality-preserving hash mapping. The hash mapping hashes nearby sensor data to the same node, and uses *k-d* tree to support range queries. DIFS [5] relies on geographic hash and *quad* tree structure for efficient index construction and range searches. DIMENSIONS [3] uses data compression and data aging to reduce redundancy caused by spatial and temporal correlation. The indexing approaches could be expensive for data storage because data can be sent far across the network and index itself can be difficult and complex to maintain.

Special path routing approaches provide a kind of information brokerage in which the producers store data not at a single node or its nearby neighbors but at nodes that follow a one-dimensional curve. The consumers travel along a set of nodes that follow another one-dimensional curve. When the two curves intersect, a consumer acquires the appropriate data. Each of these curves is drawn from a function of, respectively, the locations of producers and consumers, and not related to the type of data that is stored or queried. Examples include double ruling [14], landmark [2], and combs-needles-haystacks [9]. The common feature is that they take geographical locations into consideration but without data rates.