Research Journal of Applied Sciences, Engineering and Technology 4(8): 1004-1009
ISSN: 2040-7467
© Maxwell Scientific Organization, 2012
Submitted: January 19, 2012 Accepted: February 17, 2012 Published: April 15, 2012

Design of Large-Scale Sensory Data Processing System Based on Cloud Computing

[1]Bing Tang and [2]Yu Wang
[1]School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan, 411201, China
[2]College of Computer and Information Engineering, Hohai University, Nanjing, Jiangsu, 210098, China

Abstract: In large-scale Wireless Sensor Networks (WSNs), with limited computing power and storage capacity of sensor nodes, there is an urgent demand of high performance sensory data processing. This study studies the interconnection of wireless sensor networks and cloud-based storage and computing infrastructure. It proposes the idea of distributed databases to store sensory data and MapReduce programming model for large-scale sensory data parallel processing. In our prototype of large-scale sensory data processing system, Hadoop Distributed File System (HDFS) and HBase are used for sensory data storage, and Hadoop MapReduce is used for data processing application execution framework. The design and implementation of this system are described in detail. The simulation of environment temperature surveillance application is used to verify the feasibility and reasonableness of the system, which also proves that it significantly improves the data processing capability of WSNs.

Key words: Cloud computing, Map Reduce, sensory data processing, wireless sensor networks

INTRODUCTION

Wireless Sensor Network (WSN) enables collaboratively monitoring, sensing and collecting the environmental or object information in real-time, and processing the information, while the results will be sent to users. Wireless sensor network becomes more and more important in engineering filed, especially as the development of emerging hot Cyber-Physical Systems (CPS) technology. It is widely used in national defense and military, environmental monitoring, large-scale structural health monitoring, mine safety measurement, electrical and mechanical equipment health monitoring and other fields. It has the characters of a large network size, high node density, strictly limited node computing power and storage capability, frequently changing in network topology (Akyildiz *et al.*, 2002).

Due to the limitation of hardware design, each sensor node in a WSN is generally a simple resource-limited independent system. It means that lower communications, limited computing power, limited storage capacity, and limited battery energy. There is a high performance network aggregation node (Sink) or base station, and data collected by ordinary nodes is passed to the Sink node in the way of self-organizing and multi-hopping. Then, the sensory data is analyzed and processed by the Sink node, or even simply processed by ordinary nodes. In current large-scale WSN, how to store and process sensory data becomes the key factors which constrain and affect the application of WSN.

Large-scale WSN demands high requirements of data storage, data analysis and processing. So, what is large-scale wireless sensor network, and how to define it? The important features of large-scale WSN include:

- Involving large number of sensor nodes, typically more that 1000
- Long-term surveillance for over one year
- Generating large volume of sensory data

For example, GreenOrbs (Liu *et al.*, 2010) is a real large-scale WSN with 1000+ nodes, which realizes all-year ecological surveillance in the forest, collecting various sensory data including temperature, humidity, illumination, and carbon dioxide titer. The deployment of large-scale and long-term WSN also requires solutions for storing and reasoning huge amount of data.

On the other hand, in a data explosion era, with the development of multimedia wireless senor network, for a high-speed and continuous data stream application, huge data will be generated in the network. Thus, targeting at this kind of data-intensive WSN, how to collect, transmit, store and manage sensory data is a great challenge to sensor nodes and the Sink node, which also plays an important role to expand WSN application area.

Furthermore, the application of WSN is diversity, for example, environment temperature surveillance, structure health monitoring, sea surface monitoring, and so on. Is it possible to share a unique storage and processing infrastructure for different WSNs? This is the key

Corresponding Author: Bing Tang, School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan, 411201, China

problem we try to solve. If it can be realized, different WSNs would share the same sensory data storage utility and the same data processing framework, which would be a great improvement.

In this study, motivated by the above requirements, we propose the cloud-based approach by building an external Cloud Data Center to provide computing service and storage service for the large amount of sensory data. We also discuss the design of this cloud-based sensory data processing system, and present the implement of a prototype system. The objective of the system is to compensate the disadvantages of sensor nodes'limited computing and storage capability. In this architecture, sensory data is transferred to cloud back-end for further storing and processing through the Sink node and data gateway. The major technical difficulties to overcome include interconnection of WSN and Cloud Data Center, application porting, programming, data storage, and so on.

LITERATURE REVIEW

Data management in WSNs: Nowadays, although many works about WSN have been done, they mainly concentrate on the technical challenges related to WSN, such as MAC protocol, routing, transport, self-organization, location algorithm, and so on. Data management in large-scale WSN and the interconnection of WSN and high performance computing, such as Grid and Cloud, also start to attract our attention.

As we mentioned before, with large-scale 1000+ nodes in WSN, data storing and processing are confronted with great challenges. In traditional data management in WSN, query-based access method is widely used. Data management technologies consist of data aggregation, data storing, data querying, and data accessing, which are also the core of WSN. In particular, there are mainly three data storing strategies,

- **Centralized storing:** In which data collected by nodes is transmitted to base station for storing, accessing and processing, while there is a large node communication overhead
- **Distributed storing and indexing:** In which data is distributed in network and a data index is built for high efficient query
- **Locally storing:** In which data is stored in sensor node, so there is a lower communication overhead, also a lower query efficiency (Balazinska *et al.*, 2007)

The current representatives of sensor network data management solution include TinyDB (Madden *et al.*, 2005) and Cougar (Demers *et al.*, 2003). Both of them treat sensory data using relation table and adopt SLQ-like query language. TinyDB focuses on the realization of the low level system framework and the optimization of the query engine, while Cougar mainly focuses on effective communication mechanisms and energy optimization to achieve efficient query mechanism, as well as inner network query processing in order to reduce communication energy consumption. Data-centric network-based model is another new approach for sensor data operation and querying (Li *et al*., 2003).

Interconnection of WSNs and HPC: MIT Magazine Innovation Technology Review proposed 10 emerging technologies which will change the world, and both Grid computing and wireless sensor network are included. Grid computing creates the virtual supercomputers by using spare computing resources geographically dispersed in Internet, and computing resources are independent computing clusters which are not within a single administrative domain. Grid computing generally can offer online computation or storage. The Open Grid Services Architecture (OGSA) aims to define a new common and standard architecture for grid-based application (Foster and Kesselman, 2003).

Several initiatives around the world are studying the interconnection of vast numbers of sensor nodes with Grid computing infrastructure. Sensor Grid (Hingne *et al.*, 2003; Lim *et al.*, 2005) is such a hybrid architecture which integrates wireless sensor networks with grid infrastructures to enable real-time sensory data collection and the sharing of computational and storage resources for sensory data processing and management. It is an enabling technology for building large-scale infrastructures, integrating heterogeneous sensor, data and computational resources deployed over a wide area, to undertake complicated surveillance tasks.

Based on the idea of Sensor Grid, Hourglass (Gaynor *et al.*, 2004) project proposes the conception of Data Collection Network (DCN) which supports the interconnection of multiple WSNs through Internet. The National Weather Study Project (NWSP) (Lim *et al.*, 2007) in Singapore is another large-scale community-based environmental initiative that aims to promote the awareness about weather patterns, global warming and population. The NWSP is set up to connect the school weather stations, so that the weather data can be automatically collected and stored in a Central Data Depository (CDD) in real-time. The weather data will be made available for query, process, visualize, and archive.

In recent years, we have witness the rapid advent of cloud computing, in which remote software or storage is delivered as a service and accessed by users using a thin client over the Internet. Since Google proposed the MapReduce programming model in 2004, it has become an important cloud computing middleware (Dean and Ghemawat, 2008). Yahoo! also proposed its open source solution-Hadoop, which implements the Hadoop Distributed file system (HDFS) and Hadoop MapReduce execution framework, and distributed scalable Hadoop database (HBase) (White, 2010). MapReduce programming mode is used for data-intensive computing, which follows two stages of Map/Reduce computation to realize simple and efficient large-scale parallel data

analysis. If we can adopt the cloud computing approach and focus on service hiring by paying for storage and computation, interconnection of WSN and cloud computing infrastructure will be a new solution to sensory data storing and processing. At present, there is not a well-known mature WSN-cloud integration solution used in a real large-scale WSN system application.

SYSTEM ARCHITECTURE

General overview: Nowadays, physical world or digital world has not been independent system any more, and there is a more and more tight relationship between them, as well as more mature technologies appear, such as sensor, RFID, video surveillance cameras, and so on. In current data expansion era, data is generated all the time in our daily life. This trend requires efficient storing and processing technology for large-scale sensory data.

In order to understand real-time environment, and realize more intelligent decision making, and carry out intelligent control, and give a feedback to the physical world according to environment information, an advanced data processing, especially large-scale sensory data parallel processing framework is needed.

In this study, we propose a cloud-enabled back-end data storing and processing system for sensory data. Interconnection of WSN and Cloud computing is quite feasible, due to that there are two opportunities:

- From data intensive processing to cloud computing
- From wireless sensor network to cyber-physical system

Form the viewpoint of service computing, there are also some advantages in that cloud service could be hired by WSN system, as a remote and back-end assistant to WSN.

In our proposed approach, we summarize the hot open research issues here, which must be solved in order to realize a full prototype system:

- Data pre-processing and collection method for discrete sensory data with deviation and diversity
- The WSN and high performance data center interconnection and bridging method
- Distributed storage model for large volume WSN sensory data
- Large-scale sensory data parallel processing framework
- Performance evaluation for WSN sensory data processing

System architecture: The objective of the proposed hybrid WSN/Cloud architecture is also to realize remote management platform for sensory data storage that leverages powerful cloud computing technologies to provide excellent data scalability, rapid visualization, and user programmable analysis. It is designed to support long-term deployments of wireless sensors network through a simple Data Management API. The detailed system architecture is shown in Fig. 1. In this figure, there are three entities in this hybrid WSN/Cloud architecture: WSNs, Cloud Data Center and Users. It is a general architecture of interconnection of several WSNs through

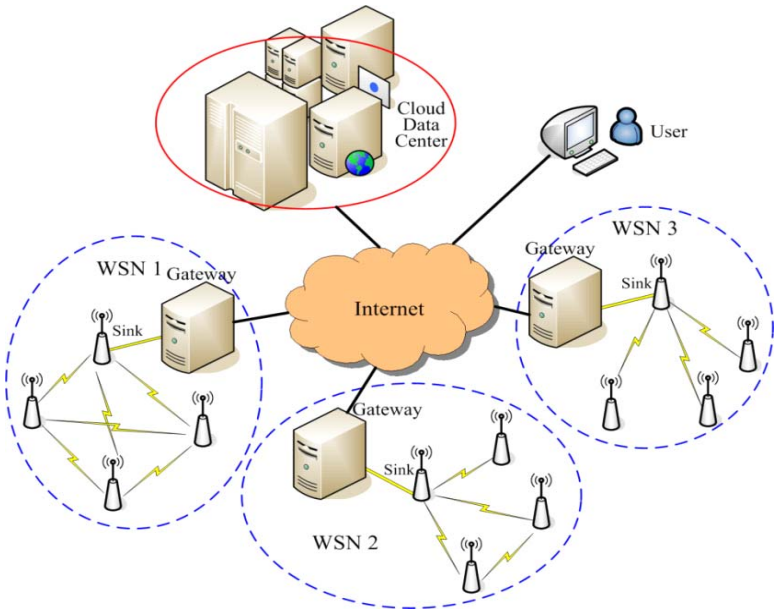

Fig. 1: Detailed system architecture of interconnection between WSN and Cloud Data Center

Internet. The server node located in the edges of blue dotted line is considered to be a data gateway, which receives data from the Sink node. Cloud-based data storing and processing center is deployed inside the red line area.

The vast amount of data collected by the sensors can be processed, analyzed, and stored using the computational and data storage service of the cloud. In this architecture, the sensory data can be efficiently shared by different users and applications under flexible usage scenarios. Each user can access a subset of the sensors, and run a specific application, and search the desired sensory data, for example, through a web-based interface.

System Implementation: Based on the statement of system architecture in previous section, we detail the modules of data storing and processing framework. Recently, there are three research highlights in cloud computing era, MapReduce, Virtualization, and XaaS, which can be used as our solution. The proposed system is just built upon scalable, fault-tolerant distributed systems-Hadoop and HBase, to facilitate data analysis. HDFS and HBase-based sensory data storage model and a flexible data processing framework based on Hadoop MapReduce can fit for a variety of applications.

Data storage model-HDFS and HBase: The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Hadoop software offers Hadoop Distributed File System (HDFS), a distributed file system that provides high-throughput access to application data. HBase is an open-source, scalable, distributed database that supports structured data storage for column-oriented large tables. It provides Bigtable-like capability on top of Hadoop and HDFS, and easy to use Java API for client access (reads and writes). It is used for hosting of very large tables atop clusters of commodity hardware and to facilitate random, realtime read/write access to your Big Data. Figure 2 shows the detail components of Hadoop-based Cloud provider, from which we know that HBase is a distributed database, and each node is managed by the master node.

Data processing framework-Hadoop MapReduce: MapReduce is a parallel programming paradigm successfully used by large Internet service providers to perform computations on massive amounts of data. After being strongly promoted by Google, it has also been implemented by the open source community through the Hadoop project. The key strength of the MapReduce model is its inherently high degree of potential parallelism. In Hadoop MapReduce framework, the computing is divided into two stages: Map and Reduce. Map processes a key/value pair to generate a set of intermediate key/value pairs, and Reduce merges all

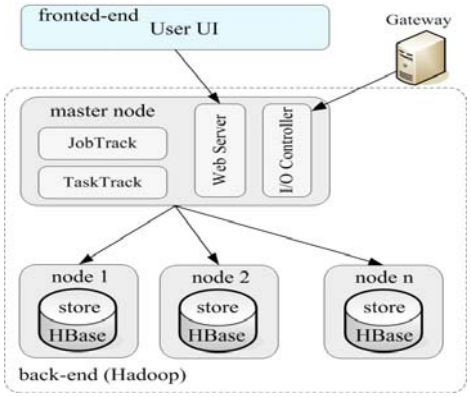

Fig. 2: The detailed components of Hadoop-based Cloud provider

intermediate values to form the final output. HBase also uses Hadoop MapReduce to process the large-sale data stored in HBase, and provides native Java API for database operation for MapReduce Job.

Interconnection of cloud and WSN: In this study, we do not address any problems of energy-efficiency communication protocol for data aggregation, sensor data fusion, data pre-processing, data filtering, and so on. We only focus on the back-end system. The interconnection is implemented by data gateway. Gateway is a client which accesses remote cloud service through Java API interface for data writes. The gateway receives data from the Sink node and then writes data into local storage as a backup, and a daemon thread is in charge of periodically writing data to Cloud Data Center. As it is shown by Fig. 2, the I/O Controller module is designed for interaction between gateway and cloud. So as to the Web Server module, it is used for provide friendly web interface for users to access sensory data, and submit query and processing request job.

SYSTEM DEPLOYMENT AND EVALUATION

Prototype deployment: The complete system is designed to scale to hundreds of sensors reporting sensing data continuously to the base station. We deployed the prototype system in student's laboratory as a demonstration. Because our system is designed to be scalable, in order to measure the advantages of proposed approach, we only built a testbed on a cluster of 8 nodes. Each node has the following configurations:

- 64-bit platform with AMD Opteron Dual-core 2.2GHz CPU and 4GB memory
- Ubuntu 8.10 Linux system
- 80 GB storage
- 100 Mb/s network interface

For the implementation of Cloud Data Center in our prototype system, the software development environments and tools are as follows: JDK 1.6, *hadoop*-0.20.2, and *hbase*-0.20.5.

In the 8 nodes Hadoop installation, one node runs the HBase master, the other 7 nodes host the HBase region servers (the servers actually storing and serving the data). After the Hadoop daemons are configured, including *NameNode*, *DataNode*, *JobTracker*, and *TaskTracker*, we start Hadoop cluster and start HBase service.

Verification and evaluation: Currently, because lack of a real environment of WSN measurement and monitor application system, we use emulation method to generate large-scale sensory data, and to emulate senor node behavior, for instance, to generate environmental temperature data. Data can also be accessed from web interface. We focus on data store, data access and processing, and the feasibility verification and a simple evaluation.

In our emulation experiment, we emulated a large-scale WSN with 1200 sensor nodes, distributed in a large area, and the sensing frequency is one time per 10 min. All the generated data are aggregated to the Sink node and gateway, and then written to HBase. In our simple verification, the data means the value of environmental temperatures. Here, we give conceptual design of two tables in HBase database, the table for storing sensor information and sensory data, respectively:

Sensor Table: { *Sensor_uuid*,
Network_id,
Sensor_id,
Sensor_position x-y}
Temperature Table: { *Sensor_uuid*,
Timestamp,
Temperature}

In our simple evaluation, we tested:

- Data storage performance, including data writes and data reads to HBase database
- Data processing performance, including searching data during a timestamp interval which belongs to a specific sensor node by SQL-like query, and executing data approximation based on cubic spline - a simple MapReduce job

In our evaluation, we run respectively two applications on two laptops which are connected to the 8 nodes Hadoop cluster we mentioned above, also with 100 Mb/s network interface. One is the emulator application which writes temperature data to the system, and the other application submits SQL-like query and data approximation job. For each write operation, 1200 sensors' environmental temperature data are stored to HBase database. For each read operation, we query the temperature value of a specific sensor, and the timestamp interval is fixed 12 h. That is to say, we read continuous 12 h's data of one sensor. To handle 12 h's discrete data (normally 72 dots in two-dimensional space of Timestamp vs. Temperature Value), cubic spline is used for approximation and making the curve more smooth.

Table 1: Performance evaluation results for data reads and data approximations

Operation times	Data read (sec.)	Data approximation (sec.)
10	2.6	1.8
50	8.2	5.9
200	33.6	21.7

The benchmark result of data reads and data approximations is shown by Table 1. Data reads and approximations are repeated for many times and we measure the time spent for these two operations. The result shows that on the assistant of high performance cloud computing, the system is reliable and quite efficient for large-scale sensory data storage and data query operations. If Cloud Data Center is not used, only relying on computing capability of sensor hardware, the Sink node, and base station, it can not be finished in such a short time for these data operations. This verification also proves that it significantly improves the data processing capability of WSN.

CONCLUSION

Based on a survey of current research status on large-scale WSN data processing and management, this paper proposes a WSN sensory data processing system using emerging cloud computing approach. High performance data processing centre is adopted as a solution to assistant the lack of sensor nodes' data storage and data processing capability. This paper also discusses the challenges of the hybrid WSN/Cloud architecture. This novel hybrid architecture enables the collection, processing, sharing, accessing and searching of large amounts of sensory data. The design and implement of the sensory data processing system are also detailed. Through a simple performance evaluation of data query and data approximation application in the prototype system, it has been proven that together with HBase-based sensory data storage, MapReduce-based data processing framework is a very promising solution to large-scale sensory data processing.

ACKNOWLEDGMENT

This study is supported by National Natural Science Foundation of China under grant no. 61103017, and Hunan University of Science and Technology Research Fund under grant no. E51097.

REFERENCES

Akyildiz, I.F., W. Su, Y. Sankarasubramaniam and E. Cayirci, 2002. Wireless sensor networks: A survey. Comput. Netw., 38(4): 393-422.

Balazinska, M., A. Deshpande, M.J. Franklin, P.B. Gibbons, J. Gray, M. Hansen, M. Liebhold, S. Nath, A. Szalay and V. Tao, 2007. Data Management in the Worldwide Sensor Web. IEEE Pervas. Comput., 6(2): 30-40.

Dean, J. and S. Ghemawat, 2008. MapReduce: Simplified data processing on large clusters. Commun. ACM, 51(1): 107-113.

Demers, A., J. Gehrke, R. Rajaraman, N. Trigoni and Y. Yao, 2003. The Cougar Project: A work-in-progress report. ACM SIGMOD Record, 32(4): 53-59.

Foster I. and C. Kesselman, 2003. The Grid 2: Blueprint for a New Computing Infrastructure. 2nd Edn., Morgan Kaufmann, USA.

Gaynor, M., S. Moulton, M. Welsh, E. LaCombe, A. Rowan and J. Wynne, 2004. Integrating wireless sensor networks with the grid. IEEE Internet Comput., 8(4): 32-39.

Hingne, V., A. Joshi, E. Houstis and J. Michopoulos, 2003. On the Grid and Sensor Networks. Proceedings of International Workshop on Grid Computing (Grid 2003), pp: 166-173.

Li, J.Z., J.B. Li and S.F. Shi, 2003. Concepts, issues and advance of sensor networks and data management of sensor networks. J. Software, 14(10): 1717-1727.

Lim,H.B., Y.M. Teo,P. Mukherjee,V.T. Lam,W.F. Wong and S. See, 2005. Sensor Grid: Integration of Wireless Sensor Networks and the Grid. Proceedings of the 30th Anniversary IEEE Conference on Local Computer Networks (LCN 2005), pp: 91-98.

Lim, H.B., K.V. Ling, W. Wang, Y. Yao, M. Iqbal, B. Li, X. Yin and T. Sharma, 2007. The National Weather Sensor Grid. Proceedings of the 5th ACM Conference on Embedded Networked Sensor Systems (SenSys 2007).

Liu, Y., G. Zhou, J. Zhao, G. Dai, X. Li, M. Gu, H. Ma, L. Mo, Y. He, J. Wang, M. Li, K. Liu, W. Dong and W. Xi, 2010. Long-term large-scale sensing in the forest: Recent advances and future directions of GreenOrbs. Front. Comput. Sci. China, 4(3): 334-338.

Madden, S.R., M.J. Franklin, J.M. Hellerstein and W. Hong, 2005. TinyDB: An acquisitional query processing system for sensor networks. ACM T. Database Syst., 30(1): 122-173.

White, T., 2010. Hadoop: The Definitive Guide. 2nd Edn., O'Reilly Media Inc., USA.