A convergent model for distributed processing of Big Sensor Data in urban engineering networks

D S Parygin[1], A G Finogeev[2], V A Kamaev[1], A A Finogeev[2], E P Gnedkova[1], A P Tyukov[1]

[1] Volgograd State Technical University, 28, Lenina Ave., Volgograd, 400131, Russia
[2] Penza State University, 40, Krasnaya Str., Penza, 440026, Russia

E-mail: dparygin@gmail.com, alexeyfinogeev@gmail.com

Abstract. The problems of development and research of a convergent model of the grid, cloud, fog and mobile computing for analytical Big Sensor Data processing are reviewed. The model is meant to create monitoring systems of spatially distributed objects of urban engineering networks and processes. The proposed approach is the convergence model of the distributed data processing organization. The fog computing model is used for the processing and aggregation of sensor data at the network nodes and/or industrial controllers. The program agents are loaded to perform computing tasks for the primary processing and data aggregation. The grid and the cloud computing models are used for integral indicators mining and accumulating. A computing cluster has a three-tier architecture, which includes the main server at the first level, a cluster of SCADA system servers at the second level, a lot of GPU video cards with the support for the Compute Unified Device Architecture at the third level. The mobile computing model is applied to visualize the results of intellectual analysis with the elements of augmented reality and geo-information technologies. The integrated indicators are transferred to the data center for accumulation in a multidimensional storage for the purpose of data mining and knowledge gaining.

1. Introduction

Automated process control systems have a widespread introduction. This fact highlights the need to collect and process a large volume of telemetry data from a multiple sensors, which are located at the monitoring objects [1]. These data is necessary for the analysis and forecasting of the condition and functioning objects, processes of technogenic and natural events [2]. The efficiency and quality of administrative decisions depend upon the following factors: continuous obtainment of information about controlled objects and processes; completeness and objectivity of the information processing; clarity of the results presentation for decision support systems [3]. However, these problems are not fully reflected in the existing papers. So the research aim is to try to eliminate this disadvantage.

2. The problems of intellectual Big Sensor Data processing

Currently, there is great scientific and practical interest in respect of wireless sensor networks (WSN) development and use [4]. Their application for remote monitoring of various objects and processes in natural, technical, environmental, medical, military and other systems is of particular importance. The tendency to substitute wired networks for wireless telecommunications for monitoring spatially distributed objects is observed. Engineering network energy, heat, water and gas, oil and gas pipelines, telecommunication networks are such objects. In engineering networks with many facilities, located in vast territories, it is not always possible to implement a full-scale collection and transmission of data in a single dispatching center for real-time processing [5]. Therefore, an actual problem of the synthesis and study of the model of collecting information based on the method of distributed computing, is realized directly to the end-points of data collection. This opportunity has come due to the spread of WSN technology. WSN units have sufficient processing power and memory to download software agents capable of solving the data processing and aggregation problem.

The convergence of the distributed computing models is proposed to solve the tasks of collecting, processing and integration of Big Sensor Data in the monitoring of distributed objects and processes. This approach includes convergence models of the grid, cloud, fog and mobile computing [6]; association of computing clusters in a single system; integration of business logic, operating platforms, data storages of server applications; unification of computing mechanism control, information security at all data processing and storage levels.

A number of research groups doing research in the field of fog and cloud computing obtain the results in terms of practical application [7]. Intel is engaged in the construction and implementation of hybrid cloud computing models. They provide energy efficiency and optimization of human resources; it is easy to create a cloud computing; there is safety information to reduce risk. The IBM Company, Google, Toshiba, Cisco, Microsoft offer solutions for analytical processing of Big Data using cloud computing models. Solutions are used to support the energy and health care sectors, to create secure data storage services, to support ubiquitous Internet technology, inter-machine communications, fog computing. One of the research areas in the Toshiba and Cisco infrastructure is the synthesis of distributed processing of data obtained from a plurality of distributed sensors based on the architecture of fog computing for the interaction of smart Internet of Things [8]. Joint development assumes that the Cisco fog computing network infrastructure is associated with Toshiba Group technologies to control points on the network, tracking and maintenance of the geographically dispersed multi-function devices.

3. The objectives of energy monitoring and distributed Big Sensor Data processing

Energy management is a power management system that allows forecasting and controlling the processes of production, transportation and use of energy to meet the needs of citizens and industrial enterprises. The aim is to achieve effectiveness in terms of costs optimization in production and transportation of energy to the end user. Energy security for risk prevention is associated with the use of automated process control systems of energy supply and consumption. A delay in power supply for the enterprises, hospitals, schools, kindergartens can lead to serious losses, social consequences, a reduced level of life safety of residents, etc. Energy management is based on the monitoring of facilities and situations to forecast economic, resource and price trends.

The energy monitoring term is used to describe the processes of monitoring and data collection with spatially distributed objects in the field of production, transport, distribution, consumption and utilization of energy resources (gas, water, electricity and heat). The aim of monitoring is measuring technological parameters for the forecast of energy production and consumption. An important task is also the identification, analysis and assessment of natural, technogenic and anthropogenic factors that affect the objects and processes. The system provides detection of abnormal and emergency situations. The purpose of the automated energy monitoring system with a converged computing model is to improve the enterprises energy efficiency, to reduce energy losses and to optimize consumption.

Energy monitoring objects are engineering communication networks. These include internal and external power supply networks of buildings, lighting networks of urban infrastructure, internal and external heat networks, water supply and sewage networks, gas supply networks, etc.

An automated monitoring system should provide the following tasks:

1. The collection, processing, analysis, storage and transmission of information on the location and state parameters of distributed objects and technological processes of generation, transportation, consumption and waste of energy and on the occurrence of abnormal situations and emergencies, etc.

2. Support for activities to ensure security engineering services, technological processes of generation, transportation, consumption and utilization, crises containment, consequences liquidation.
3. Synthesis and analysis of forecast models for the integrated assessment of freelance, emergency and crisis situations in relation to the monitoring of distributed objects and processes.
4. Forecasts of threats to citizens, utilities, facilities and technological processes depending on changes in the state because of natural, technological and anthropogenic factors.
5. Evaluation of different situations when monitoring objects and scenario analyzing.
6. Maintaining operational databases, cloud storage and data marts for the storage and maintenance of the potential availability of sensor data, and aggregating monitoring results, the synthesis variants of administrative decisions on the protection of objects, subjects and processes.
7. Supply of information resources for the administration, protection of resources against unauthorized access and information threats.
8. Formation of a heterogeneous transport environment, and a single information space of monitoring systems based on standardization and compatibility of information, software and hardware.
9. Information support of agreements in the process of engineering communications monitoring, processes of generation, transportation, energy consumption and utilization.

The result is the development and implementation of measures aimed at ensuring optimal and trouble-free operation of engineering services networks in urban energy systems. Activities are provided by: improving the efficiency of control services; energy saving and minimization of energy losses; optimization of energy consumption; implementation of data collection and processing technologies; prevention of emergency situations; rapid response to the situation and the disaster; maintenance and repair works in the engineering systems, etc.

4. Models and methods of convergent distributed computing

Today cloud services ensure an efficient use of applications by limiting capital investments and reducing the cost of corporate information systems ownership. Four models of distributed computing can be distinguished: (i) the grid computing, (ii) cloud computing, (iii) fog computing and (iv) mobile computing. The grid calculations are based on the architecture of computer networks of the individual compute nodes. The computing process provides for a distribution of separate parts in order to free network computing resources. This approach is used for the tasks too complex for a single node.

Cloud computing is not only the allocation of tasks on the network nodes of computing resources. This model is used for the ubiquitous network access to a common pool of configurable resources (software, server, information, platform, etc.) at any time. The user applies the technology of a “thin” client as a means of access to applications, platforms and data, and the entire infrastructure of the information system is located at a provider of cloud services.

Fog computing is a virtual platform of distributed computing and data storage services on end-terminal devices and network services for data transmission, storage and processing [9]. Computation is performed by terminal devices with limited computing and energy resources.

The reasons for the development of a converged model of distributed computing are the problems of information processing:

1. The opacity information, which is associated with the close format data information systems from various developers. This is unacceptable for servicing critical and potentially dangerous objects.
2. Mismatch of data and protocols that is associated with the use of manufacturers proprietary systems for collecting and telemetry data processing.
3. Duplication and synchronization of information.
4. Mismatch of data streams, associated with the use of different network technologies.
5. Qualitative data accounting to monitor energy supply. The problem occurs because of differences in the departmental affiliation of electricity, heat, water and gas enterprises. To solve the problems, the cloud storage and fog computing technologies are used. The problem of duplication data in the database of the different services is solved by using a single data store. Fog computing technology allows to perform data processing in terms of their format unification.

5. A wireless sensor network as a fog computing platform

The first level convergent computing model is expedient to implement a fog computing platform through the software agents that are integrated into the wireless sensor network units and/or industrial controllers. To implement the platform WSN, based on ZigBee technology, is used.

Terminal nodes distributed along the periphery of the network and information processing are performed in real time during data collection. According to the model of fog computing sensors, nodes are not only used to collect data from sensors and automation devices, but also for their structuring, aggregation, encryption and transmission to the control centre. Fog processing is meant to calculate the aggregated energy indicators directly in the network nodes. Indications are forwarded for integration into the next level model of the data collection/processing. The ultimate goal is to extract information and intelligent analysis from clusters of the virtualized grid system.

Telecommunication environment for united sensor nodes distributed over a large area leads to the risk of information security being threatened by industrial viruses. It can be remotely loaded into controllers and WSN sensor nodes; then it can migrate to industrial networks and perform certain destructive actions. A similar approach is used to create intelligent software agents that are loaded into remote terminal devices and to solve data storage problems, to calculate aggregated indicators, etc.

Transferring operations at the sensor nodes associated with the primary sensor data processing are:

1. Reducing workload of low-speed data channels in sensor networks.
2. Increasing the network autonomy by reducing the number of transmissions and traffic.
3. Reducing the load on the server applications.
4. Decreasing queue lengths in data processing in the computing cluster nodes.
5. Decreasing the amount of data in a central repository.
6. Sensor data processing in real time.
7. Monitoring sensors' work directly on the terminal nodes.

6. The hyper graph model of fog Big Sensor Data processing

A converged computing system can be represented as a hypergraph model with two subsets of vertices and edges; an expanded set of features

$$G = (V(V^{id}_{\{x,y\}}, V^{pa}_{\kappa}),\ U(U_{const}^{id}, U_{var}),\ P). \quad (1)$$

A subset of vertices $V^{id}_{\{x,y\}}$ describes the network nodes with weight (*id*) attributes, characterizing the features of controllers and spatially distributed nodes coordinates $\{x,y\}$. A subset of vertices V^{pa}_{κ} identifies software agents with attributes (*pa*), characterizing parameters of the modules and attributes that define computational functions. The U_{const}^{id} constant incidence describes the data channels with weighting *(id)* attributes. A subset of hyper edges of U_{var} variable incidence describes the virtual routes for migration downloadable software agents on hosts and controllers. "*P*" is a binary predicate that determines the incidence of vertices and hyper edges.

A subset of vertices $V^{id}_{\{x,y\}}$ defines different types of devices, including processors sensor nodes and the controllers control equipment, accounting and control of energy resources, as well as different operating platforms and software firmware. Therefore, clusters of similar vertices, grouped by the processing power and platforms, are allocated in this subset. It allows downloading a certain class of agents, which is a subset of vertices V^{pa} also divided into clusters. Hyper edges of constant incidence U_{const} $(a_1, \ldots, a_n)$, where $n=const$, are divided into clusters, combining the particular segments of the network technology. Hyper edges of U_{var} combine the modeling software agents of subset vertices with the modeling specific types of subset vertices of controllers and sensor nodes.

There are features of the hypergraph model: a set of hyper edges of U_{var} $(b_1,\ldots,b_m)$ is dynamically changed in real time; vertices have geospatial tagging; the model allows to perform a space-time analysis of the software agents migration. The relational model is used for the semantic description of the convergent system. The hyper edge of U_{const} corresponding to the static entry and the hyper edge of U_{var} is a dynamically changing record.

Let the sensor data be collected and processed on the *i* node of the sensor network, and a software agent migrates through the *j* virtual route to the energy costs of its transfer q_i within the specified time interval. Each node has a reserve of e_i energy for battery life, and it spends energy e_j of data processing j of the software agent. The battery life of the sensor node may be defined as:

$$t_i = e_i/(\sum_{i\in N}(q_i) * \sum_{j\in M}(e_j)), \qquad (2)$$

where N is the number of agents migration routes, M is the number of data processing nodes.

7. The system architecture of convergent Big Sensor Data processing

The system architecture implementing a converged model of distributed computing may include five hardware and software levels:

1. The set of sensor nodes associated with industrial controllers that implement fog computing.
2. Coordinators of clustered sensors segments; cellular modems; the repeating router, etc.
3. The grid cluster and cloud computing.
4. Storage of sensor data, aggregated indicators and monitoring results.
5. The set of user devices for organizing a universal access to information resources of the system.

Hypervisor management is used by software agents. It consolidates computing resources in a multiprocessor system for distributed processing of Big Sensor Data. Software agents operate in OS sensor nodes and interact with the data acquisition modules, other agents and brokers. In this model, a computing agent is a software template for parallel processing. The intelligent agent responds to requests, decides on the selection of data processing functions, clone and migrate to other nodes.

A feature of the agents is the realization of a behavior. The behavior is determined by the mathematical function which implements the steps of processing Big Sensor Data.

The agent’s migration is the ability to duplicate itself and distribute copies to other nodes in the network. Security agents can be provided by migration prohibition, automatic firmware and OS updates inability; using means of digital signatures and key management procedures, etc.

Model intellectual brokers are offered for agent’s interaction with server applications at the data center. Broker is an agent that runs on routers and realizes data storage, protection and transmission.

Collecting cloud computing results entered from sensor segments of the ZigBee network is performed by using the broker Message Query Telemetry Transport (MQTT), loaded into a computing network gateway cluster. The broker’s functions are processing of sensor data and aggregates entered by the coordinator; conversion frames with this information are to be integrated into the data storage; data encryption; “sliding” window algorithm support for reliable transmission, etc.

8. Software tools for the intellectual Big Sensor Data processing

The distributed computing Erlang programming language is used for the development of the converged computing system software, including software brokers, software agents and server applications [10]. The language includes a means of generating parallel processes and ensures their interaction through the exchange of asynchronous messages in accordance with the multi-agent model. Failsafe is achieved by using insulated from each other lightweight processes related messaging and output signals. The Erlang program is translated into bytecode executable virtual machines on a network nodes. The system supports "hot swapping" of code brokers and agents. The use of lightweight processes in accordance with the model allows agents to perform simultaneous multiple processes on distributed nodes with limited computing resources. The processes are isolated from each other, but can be set as asynchronous messaging for the TCP/IP protocol. The communication between processes and nodes is performed using SSL and steganographic key management schemes. To create server applications in the Erlang language, the set of behaviors framework Open Telecom Platform (OTP) is used. The OTP-behaviour is divided into working and monitoring processes.

To improve the efficiency of data processing and storage units in a multidimensional SQL, there is combination of an industrial storage with a non-relational data storage system. Along with Oracle Database, a distributed non-relational Cassandra system is used for caching sections of the multidimensional storage. This improves the data sampling rate, fault tolerance and scalability.

Data mining agents also operate in a server GRID cluster. They are developed on the platform of Java Enterprise Edition (J2EE) using the Spring framework multilayer technology platform and

Object-Relational Mapping (ORM) Hibernate. Object-relational adapter Hibernate is used for querying flexibility and transparency to the storage system through Cassandra.

The server platform is built using the JBoss Application Server applications. The communication between the “thin” clients and servers is performed on the basis of HTTPS and AMF (Adobe Media Format) using the Adobe Flex and ActionScript technology platform. The three-level architecture computing cluster includes a central server, server’s GRID, and a lot of GPU video cards.

9. Conclusion

The research towards implementing the converged computing model led to the following conclusions.

1. A convergent approach to distributed computing is the convergence of distributed data processing technologies. The model is designed for the collection, processing and integration of Big Sensor Data obtained in the monitoring and control of spatially distributed objects and processes.
2. A convergent model of distributed computing includes four levels of processing and storage of sensor data. The first is the level of fog computing. At the next level, sensor data and aggregates are integrated in the multidimensional cloud storage. The third level of data processing is implemented in the server cluster. The fourth level is implemented on mobile devices where operating agents to retrieve and visualize the intelligent analysis results in the use of geo-information technologies.
3. Information interaction of the agents of fog computing along with cloud storage and database and the server applications are provided by brokers through the intellectual MQTT protocol. The hypervisor network is used for managing agents and brokers. It consolidates distributed resources in a multiprocessor complex. The functionality of agents and brokers is defined as a mathematical function that determines the action of sensor data processing and the selection of behaviors in emerging events.
4. The benefits of large sensor data processing, based on the convergent model of distributed computing, decrease the load of the server cluster, reducing the traffic volume in sensor networks, increasing the battery life of a network and its components, monitoring in real time, etc.

Acknowledgments

The reported study was funded by RFBR according to research projects No. 16-07-00388_a, No. 16-37-50017_mol_nr, No. 16-07-00353_a, No. 16-07-00031_a, No. 15-57-54033_Vet_a, No. 16-37-00387_mol_a, No. 15-07-01720_a and the Russian Ministry of Education within the framework of the base part (project 2586 of task N 2014/16).

References

[1] Sidorov V V 2016 *Proc. 3d Int. Conf. on Information technologies in science, management, social services and medicine* (Tomsk: TPU) pp 740–742

[2] Bolshakova O N and Chusavitina G N 2016 *Proc. 3d Int. Conf. on Information technologies in science, management, social services and medicine* (Tomsk: TPU) pp 122–126

[3] Sadovnikova N, Parygin D, Kalinkina M, Sanzhapov B and Trieu Ni Ni 2015 *Communications in Computer and Information Science* (Springer IPS) **535** 488–499

[4] Aleisa E 2013 *Procedia Computer Science* **19** 232–239

[5] Finogeev A, Fionova L, Finogeev A, Nefedova I, Finogeev E, Vinh T Q and Kamaev V 2015 *Communications in Computer and Information Science* (Springer IPS) **535** 474–487

[6] Nurlan Z and Zhukabayeva T K 2016 *Proc. 3d Int. Conf. on Information technologies in science, management, social services and medicine* (Tomsk: TPU) pp 703–706

[7] Arunkumar G and Venkataraman N 2015 *Procedia Computer Science* **50** 554–559

[8] Hersent O, Boswarthick D and Elloumi O 2012 *The Internet of Things: Key Applications and Protocols* (New York: Wiley) p 370

[9] Stojmenovic I and Wen Sh 2014 *Proc. 2014 Federated Conf. on Computer Sci. and Information Systems (ACSIS)* (Warsaw) **2** 1–8

[10] Nyström J H, Trinder P W and King D J 2008 *Concurrency and Computation: Practice and Experience* **20(8)** 941–968