WORKFLOW MODELLING IN GRID SYSTEM FOR SATELLITE DATA PROCESSING

Andrii Shelestov

Abstract: This paper focuses on a problem of Grid system decomposition by developing its object model. Unified Modelling Language (UML) is used as a formalization tool. This approach is motivated by the complexity of the system being analysed and the need for simulation model design.

Keywords: Grid system, satellite data processing, modelling.

Introduction

In the past decades a great attention has been paid to the development of Grid technologies. Description of the state-of-the-art in the Grid system development can be founded in [1, 2, 3, 4, 5]. Since Grid system is a complex distributed system that enables solution of complex problems in different domains, the development of such systems requires the use of system analysis approach. It is especially important to Grid systems for environment monitoring where heterogeneous resources are used: satellite and in-situ data, modelling data, etc. Such Grid system is not just a computational system, or it is intended for data management either. It represents a set of virtual organizations that jointly solve complex problems using data from geographically distributed archives with given level of reliability and security. The examples of such initiatives include GMES (Global Monitoring for Environment and Security) [6], GEOSS (Global Earth Observation System of Systems) [7] and WAG (Wide Area Grid) [1], where Ukraine is among the participants. Analysis of such complex systems requires approaches that involve system decomposition and structural and functional analysis of system components to provide the further optimization and efficient management [8].

In this paper we carry out Grid system decomposition by developing its object model. For this purpose Unified Modelling Language (UML) is applied.

Properties of Grid systems for satellite data processing

Grid system integrates computational, informational and other kinds of resources that are managed by geographically distributed organizations. Traditionally such systems are used for complex computational problems solving that require the use of high-performance computing, or problems that require processing of high volumes of data. The example include EGEE (http://www.eu-egee.org) project that is aimed at solving problems in high-energy physics [9], gravitational waves [10], astronomy [11], and bio-informatics [11]. Such Grid systems are considered as computational Grids.

Recent years much attention has been paid to the development of Grid system for environmental monitoring with the use of satellite data [2]. In such systems Grid infrastructure can be used not only for providing high-performance computations, but also for efficient data management. The Earth observation (EO) data should be processed, catalogued, and archived [13]. For example, GOME instrument onboard Envisat satellite generates nearly 400 Tb data per year [14]. EUMETCast system for environmental data dissemination that is a part of global GEONETCast system of GEOSS [15] enables acquisition of more than 50 Tb of processed and unprocessed information per year. Moreover, satellite data are processed not by the single application with monolithic code, but by distributed applications. This process can be viewed as complex workflow that is composed of many tasks: geometric and radiometric calibration, filtration, reprojection, composites creation, classification, products generation, post-processing, visualization, etc. [16]. For example, calibration and mosaic composition of 80 images generated by ASAR instrument onboard Envisat satellite takes 3 days on 10 workstations of Earth Science GRID on Demand that is being developed in ESA and ESRIN.

Thus, complexity of Grid systems for environmental monitoring and decision support using satellite data is of particular interest from system analysis point of view. The investigation of specifics and peculiarities of such systems is primary objective of DGREE project [17] initiated within EGEE-II project. The reported results provide the following requirements to the Grid system for satellite data processing:

— Security requirements: satellite data and corresponding software are distributed according to specialized licenses. So, observance of license agreements is one of the main issues of such systems.

— Reliability: such systems should operate in operational mode and should provide required reliability and QoS of the results.

— Standardization: applied problems should be solved according to standardized, verified and validated methods. Interoperability issues are also of a great concern.

These requirements provide flexibility during workflow execution, and necessity to efficient management of resources on physical and logical layers. One of the important issues is structural and functional analysis based on system modelling.

Grid System Modelling

The task of Grid system modelling is very important and motivated by the following goals. Since the development of Grid system requires large amounts of financial resources to be spent, the modelling will allow the developer to design optimal architecture of a such distributed system. Another issue corresponds to the modelling of the existing system. The development of system workload model will allow one to reveal system bottleneck, to estimate system capacity and to plan future trends of its extension. And at last, system modelling is a part of task of system performance forecasting and development of scheduling techniques for efficient resource management. By resource management we mean estimation of optimal computational resources parameters to meet the requirements of applied problems being solved, resource discovery, reservation, scheduling and monitoring. To develop optimal methods for resource management one need to run many experiments with established parameters of the system and established external conditions. Such experiments in real systems are almost impractical due to the following reasons:

— computational resources are managed by different organizations that complicates experiment;

— user requirements and properties of system resources are evolving through the time that makes it difficulty to repeat the experiment under the same conditions;

— development of Grid infrastructure to the specific experiment is time and cost-consuming.

All these factors make the modelling the only practical approach to the analysis of Grid infrastructure, its internal properties and analysis of influence of external conditions.

Approaches to Grid System Modelling

Different approaches to Grid system modelling can be used among which we can consider analytical or statistical models [18]. As a rule, analytical expression can be used to describe the properties of the system under specific assumptions such as independence of parameters, linearity, instantaneity of state transitions, etc. If these assumptions apply, the model will be in good correspondence with real object. Otherwise, there will be considerable difference between the model and real object.

That is why, for the modeling of complex distributed Grid systems simulation modeling is applied. Simulation entails the functioning of the Grid system being analyzed by integrating its elements in the single structure and imitating its interaction. Advantages of simulation modeling approach are its generality, the possibility to simulate systems of any complexity, possibility to acquire new data about properties of the system. All these factors enable detailed analysis of the system and its components. In this paper simulation modeling is applied to the Grid system intended for satellite data processing.

Classification of tasks in Grid system for satellite data processing

One of the main issues in Grid system modelling is task description and modelling. Data about tasks represent as a rule inputs to Grid system models, as well as system load for estimation and forecasting of system productivity. That is why adequate description of tasks in Grid systems will enable effective modelling of Grid system and its load. In the framework of task 2.6 tasks in Grid system for satellite data processing were classified on *Data Transfer Task* (DTT) and *Computational Task* (CT). These types of tasks represents “building blocks” for more complex tasks for environmental monitoring and decision support.

Data Transfer Task are characterized by the volume of transferred information which provides the following requirements to the system: bandwidth connection (to Internet or local network), data storages (hard drives or magnetic strips) I/O speed. Data sources, frequency and QoS should also be taken into account.

DTTs are represented by the following set of parameters:

- task identifier;
- frequency (regularly or by user query);
- input and output data volumes;
- data source (e.g. Internet, local network, hard drive, magnetic strips, etc.).

Computational Tasks represent the unit of program that carry out data processing or computations. CT can be run on either single processor or in parallel mode on different processors.

CTs are represented by the following set of parameters:

- task identifier;
- frequency (regularly or by user query);
- complexity of problem (computational complexity, memory requirements, software and hardware requirements, etc.);
- parallelism by code or data;
- number of processes required.

Another type of tasks that should be mentioned is the task of control – database search or other control instructions. But these tasks can be viewed as CT ones.

Formal task description

In order to begin with formal description of tasks with definition of Data type structure. Data can be represented by the following setoff parameters:

$$Data = \{ID, V, DS, Sec\}, \quad (1)$$

where ID — task identifier, V — data volume (in Mb), DS — data source: local discs, data storage, resource in local system or Internet. DS can be one of the following values:

DS = {local disc, Internet, data storage, local network}.

In (1) Seq describes level of required security.

DTT can be described by:

$$DTT = \{ID, Freq, I/O: Data\}, \quad (2)$$

where ID — task identifier, Freq — frequency of task completion:

$$Freq = \{cycle, request\}. \quad (3)$$

If the task is carried out periodically then Freq = cycle, and determines the period of time the task is being completed (e.g. in minutes). If the task is initiated by the query of user or other system then Freq = request, and the query time should be given by some distribution.

CT can be described by the following set:

$$CT = \{ID, Freq, C, Par\}. \quad (4)$$

In Eq. (4): ID — task identifier, Freq — frequency according to (3), C — task complexity described by the following equation

$$C=\{CC, Size, Op\}, \quad (5)$$

where CC — computational complexity, Size — required memory, Op — additional requirements (e.g. requirements to software or hardware).

In Eq. (4) Par defines parallelism of the task: the type of parallelism (by code or data), and number of required processors NumProcess.

Based on the described analysis UML class diagram for tasks of Grid system for satellite data processing was developed (Fig. 1).

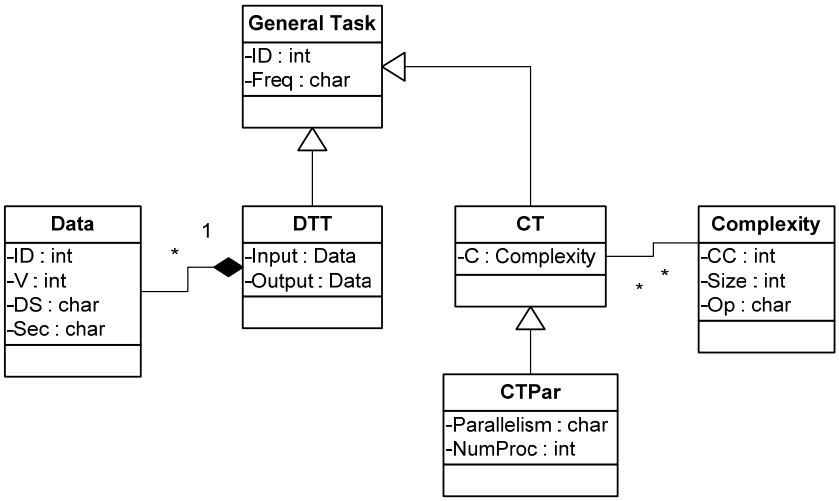

Fig. 1. Class diagram for tasks running in Grid system for satellite data processing

Description of complex tasks in Grid system for satellite data processing

The set of tasks that compose a workflow – a job – can be described by the directed acyclic graph (DAG). The nodes of the DAG represent elementary tasks – CT or DTT. Edges correspond to interdependency between tasks. Thus, a job is described by the following experession:

$$Job = \{S, W, QoS\},$$

where $S = CT \cup DTT$ — the set of basic tasks: CT or DTT, $W \subset SxS$ — set of pairs (s_i, s_j) with edge that directed from s_i to s_j, QoS — quality of service (for example, TTS — time-to-schedule time, maximum number of repeated completions).

Example of object model for the problem of biodiversity estimation using remote sensing data from space

In the framework of innovation project Space Research Institute NASU-NSAU and Centre for Aerospace Research of the Earth NASU have developed information service for estimation biodiversity in Near Black Sea region using remote sensing data from space [19]. Biodiversity refers to the variation of taxonomic life forms within a given ecosystem, biome or for the entire Earth. Species biodiversity is characterised by two criteria: the total number of species and distribution between species [20].

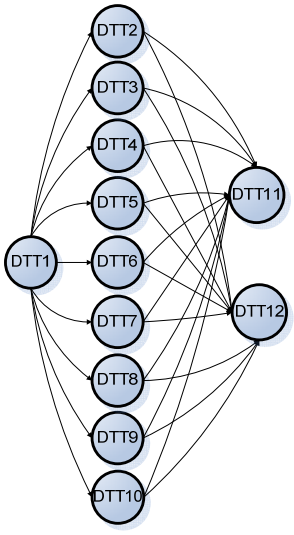

Fig. 2. DAG for data acquisition process

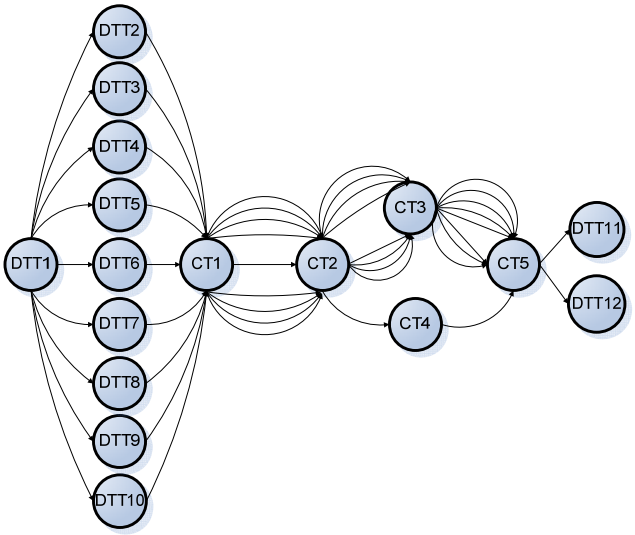

Fig. 3. DAG for data processing step

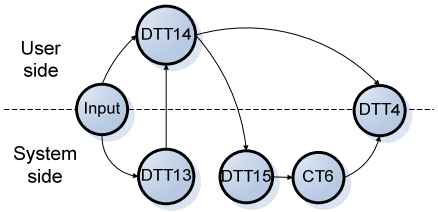

Fig. 4. DAG for the task of visualization of processing results

The workflow for biodiversity estimation using satellite data acquired by MODIS instrument aboard Terra satellite is organised as follows:
— data acquisition;
— data processing;
— visualization of the results.

The detailed description of these steps and corresponding sequence diagrams for this task are given in [21]. In this paper we focus on the presentation of workflow in the form of direct acyclic graph (DAG). Using the mentioned above approach to the decomposition of complex tasks into simple one, we represent the complex workflow for biodiversity estimation as DAGs (Fig. 2-4).

Corresponding notations to the figures 2-4 are given in table 1.

Table 1. Notations to the DAGs for biodiversity estimation task

Title	Description	Tasks structure
DTT1	Data transfer initialization	
DTT2	Data transfer (DT) MOD03A2	{cycle/24 h, I/O: 30 Mb, Internet}
DTT3	DT MOD05L2	{cycle/12 h, I/O: 6 Mb, Internet}
DTT4	DT MOD11A1	{cycle/12 h, I/O: 20 Mb, Internet}
DTT5	DT MOD12Q1	{cycle/0,5 year, I/O: 3 Mb, Internet}
DTT6	DT MOD13Q1	{cycle/16 days, I/O: 30 Mb, Internet}
DTT7	DT MOD15A2	{cycle/16 days, I/O: 5 Mb, Internet}
DTT8	DT MOD17A3	{cycle/8 days, I/O: 7 Mb, Internet}
DTT9	DT AE_Land3	{cycle/1 months, I/O: 30 Mb, Internet}
DTT10	DT SRTM DEM	{request, I/O: 1,5Gb, Internet}
DTT11	Data saving (in storage)	
DTT12	Updating data (in index service)	
CT1	Data reprojection	CT = {request, C, DP} C={O(N*M), 1,61 Gb}, where DP — data parallelism; N, M — image sizes.
CT2	Data scaling	CT = {request, C, DP} C={O(N*M), 1,61 Gb}
CT3	Composite creation	CT = {request, C, DP} C={O(N*M*t), 101*t Mb}, where t — averaging period.
CT4	Solar irradiation estimation using digigtal elevation model	CT = {request, C, PD} C={O(N*M), 1,5 Gb}
CT5	Biodiversity estimation	
DTT13	Obtaining parameters of processed data	{request, I/O: 100Kb, storage}
DTT14	Selecting parameters to be visualized	{request, I/O: 100Кб, Internet}
DTT3	Obtaining data for layers generation	{request, I/O: 1Mb, storage}
CT6	Layer generation by map server	CT = {request, C, DP} C={O(N*M), 1Mb}
DTT4	Transferring data to client browser	{request, I/O: 1Mb, Internet}

The presentation of complex tasks in the form of DAG is more suitable for scheduling and allows one to identify possibilities for parallelization.

Conclusions

In this paper we described requirements to the Grid systems aimed at satellite data processing for applied problems solving. We provided classification and description of different types of tasks executed in such Grid systems. We proposed an object model of complex task that is composed of a set of simple tasks and presented it in the form of directed acyclic graph. Such approach enables automatic creation workflows that need to be executed in Grid system. We applied our approach to task of biodiversity estimation using remote sensing data from space. Future works will be directed to the development of model of system resources and extension of software tools for simulation of Grid systems.

Acknowledgement

This research is supported by the Science and Technology Center in Ukraine (STCU) and the National Academy of Sciences of Ukraine (NASU) within project “GRID technologies for environmental monitoring using satellite data”, no. 3872.; and INTAS-CNES-NSAU project “Data Fusion Grid Infrastructure”, Ref. No 06-1000024-9154.

Bibliography

1. Shelestov A.Ju., Kussul N.N., Skakun S.V. Grid-technologies in monitoring systems based on satellite data // J. of Automation and Control. — 2006. — № 1-2. — P. 259-270. (in Russian)
2. Shelestov A.Yu., Kravchenko A.N., Korbakov Ì.B., Kussul N.N., Skakun S.V., Rudakova A.I., Illin N.I., Tyutyunnik L.I. Grid-technology of Implementation of GEOSS Ukrainian Segment // J. of Communications (special edition). – 2006. - P. 106-125. (in Russian)
3. Zagorodniy A.G., Zinoviev G.M., Martynov E.S., Svistunov S.Y., Shadura V.M. Grid – new infroamtion and coputation technology for science // Bulletin of NASU. — 2005. — № 6. — P. 17-25. (in Ukrainian)
4. Zgurovsky M.Z. Development of educational and research segment of information society in Ukraine // Системні дослідження та інформаційні технології — 2006. — № 1. — С. 7-17.
5. Shelestov A.Ju., Korbakov M.B., Lobunets O.G. Implementation of Grid-infrastructure for satellite data processing // J. Problems of Programming. – 2006. - №2-3. – P. 94-101. ISSN1727-4907 (in Russian)
6. Building a European information capacity for environment and security. A contribution to the initial period of the GMES Action Plan (2002-2003) // Office for Official Publications of the European Communities (Luxembourg). — 2004. — 238 p.
7. Global Earth Observation System of Systems (GEOSS), 10-Year Implementation Plan Reference Document // ESA Publication Division, Netherlands, 2005. — 209 p.
8. Zgurovsky M.Z., Pankratova N.D. System analysis: problems, methodology, applications, Kyiv, Naukova Dumka, 2005, 743 pages.
9. Holtman K. CMS Requirements for the Grid // Proceedings of the International Conference on Computing in High Energy and Nuclear Physics (CHEP2001). — 2001.
10. Deelman E., Blackburn K., et al. GriPhyN and LIGO, Building a Virtual Data Grid for Gravitational Wave Scientists // Presented at 11th Intl. Symposium on High Performance Distributed Computing. — 2002.
11. Annis J., Y. Zhao et al. Applying Chimera Virtual Data Concepts to Cluster Finding in the Sloan Sky Survey // Technical Report GriPhyN-2002-05, 2002.
12. Peltier, S.T., et al. The Telescience Portal for Advanced Tomography Applications // Journal of Parallel and Distributed Computing: Computational Grid. — 2002. — 63(5). — P. 539-550.
13. Fusco L. Earth Science GRID on Demand // CEOS WGISS-21 GRID Task Team Meeting. — Budapest, May. — 2006.
14. Fusco L., Goncalves P., Linford J., Fulcoli M., Terracina A., D’Acunzo G. Putting Earth-Observation on the Grid // ESA Bulletin. — 2003. — 114. — P. 86-91.
15. GEONETCast, http://www.earthobservations.org/progress/GEONETCast.html.
16. Reis W.G. Remote Sensing Basics, Moscow: Technosphera, 2006, 336 pages. (in Russian)
17. Dissemination and Exploitation of GRids in Earth sciencE. — http://www.eu-degree.eu.

18. Popkov Yu.S. Macro-systems and Grid technologies: modelling dynamical stochastic networks // J. of Automation and Information Science. — 2003.— Num. 3.— P. 10-20. (in Russian)
19. Kussul N., Popov M., Shelestov A., Stankevich S., Korbakov M., Kravchenko O., Kozlova A. Information service for biodiversity assessment in Pre-Black Sea region in the framework of development of Ukrainian segment of GEOSS // J. Science and Innovations. — 2007. — Vol. 3, Num.6. — P. 17-29. (in Ukrainian)
20. Odum Yu. Ecology.- Vol.2.– Moscow: Mir, 1986.– 376 p. (in Russian)
21. Shelestov A.Yu. An object task model in satellite data processing Grid system// Transactions of Donetsk National Technical University. — 2007. — Issue 8(120). — P. 317-330.

Authors' Information

Andrii Yu. Shelestov – PhD, Senior Researcher, Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, Kyiv-187, 03680 Ukraine, e-mail: inform@ikd.kiev.ua.

LEONTIEF MODEL ANALYSIS WITH FUZZY PARAMETERS BY BASIC MATRIXES METHOD

Vladimir Kudin, Grigoriy Kudin, Alexey Voloshin

Abstract: The basic matrixes method is suggested for the Leontief model analysis (LM) with some of its components indistinctly given. LM can be construed as a forecast task of product's expenses-output on the basis of the known statistic information at indistinctly given several elements' meanings of technological matrix, restriction vector and variables' limits. Elements of technological matrix, right parts of restriction vector LM can occur as functions of some arguments. In this case the task's dynamic analog occurs. LM essential complication lies in inclusion of variables restriction and criterion function in it.

Keywords: Leontief model, quantitative and qualitative analysis, fuzzy set, basic matrix, membership function.

Introduction

Mathematical apparatus of fuzzy sets is the way of indefinite parameters assigning, the values of which are unknown until the moment of decision-making. One of the mechanisms of vagueness removal in parameters assigning at model construction is the presence in the outline the decision-making person (DMP). DMP is aimed in workmanlike manner to determine the model's structure, to indicate the mechanism of vagueness removal at its formation [Orlovskij, 1981]. LM essential complication (LM) [Leontief, 1972], [Hass, 1961] is the inclusion of restrictions on variables' meanings (values) [Orlovskij, 1981]. One of the LM peculiarities is the inclusion of mathematical problems analysis series of linear systems as systems of linear algebraic equation (SLAE) with the quadratic nondegenerate matrix of restrictions, linear algebraic inequalities (SLAI), with the corresponding matrix of restrictions and also the tasks of linear programming (TLP) [Voloshin, 1993], [Vojnalovich, 1987], [Vojnalovich, 1988], [Kudin, 2002]. Realization of model's qualitative analysis [Orlovskij, 1981] predetermines as well the inclusion of quantitative analysis of its structural elements' consistency [Voloshin, 1993], [Vojnalovich, 1987], [Vojnalovich, 1988], [Kudin, 2002]. We can differentiate the following stages of analysis:

- testing of mathematical and computer-assisted non-degeneracy of restrictions matrix, determination of its rank's value;
- directing correction of restrictions matrix's rank's value with the means of changing its single elements (in case of necessity);
- revelation of LM common features itself and the restrictions on variables – solubility (insolubilities);