Институт проблем информации
Федеральный исследовательский центр "Информатика и Управление"
Российской Академии Наук
Национальный Исследовательский Ядерный Университет «МИФИ» (НИЯУ МИФИ)
Обнинский Институт Атомной Энергетики (ИАТЭ)
Московская секция ACM SIGMOD
Российский фонд фундаментальных исследований

Аналитика и управление данными в областях с интенсивным использованием данных

XVII Международная конференция
DAMDID / RCDL′2015

Обнинск, 13 – 16 октября 2015 года

Труды конференции

Institute of Informatics Problems
Federal Research Center “Computer Science and Control”
of the Russian Academy of Sciences
Obninsk Institute for Nuclear Power Engineering, branch of MEPHI
Moscow Chapter of ACM SIGMOD
Russian Foundation for Basic Research

Data Analytics and Management in Data Intensive Domains

XVII International Conference DAMDID/RCDL’2015

October 13 – 16, 2015, Obninsk, Russia

Proceedings of the Conference

Large-scale data analysis of random growth structures: software and hardware platform.

© Anton Yu. Menshutin © Lev N. Shchur
Science Center in Chernogolovka, 142432 Chernogolovka, Russia

may@itp.ac.ru shchur@chg.ru

Abstract

We develop integrated software and hardware platform for simulation of growth structures. This platform addresses problems of generation, analysis and visualization of random growth structures in a user-friendly way. Our intent was to hide the complexity of the simulation by means of an underlying middle-ware which includes Linux OS in a clustered environment with parallel job batch system, distributed network file system for bulk data storage, and web interface for GUI. We discuss various issues when implementing such an integrated platform, and debate on the future potential of such systems.

1 Introduction

Computational analysis is an indispensable part of the most modern-day scientific research. Research outcome strongly depends on the ability of the research team to process and analyze the data. Data analysis can be complicated. For instance, consider finding the numerical solution of a system of equations that describe some system behavior, or finding the numerical simulation of a system based on first-principles laws. While the availability of computational power provides ample opportunities for researchers, it also imposes an extra requirement on them. In addition to their respective fields of study, researchers must also have a deep knowledge in the field of computer science.

That dependence on computer analysis is ubiquitous across all fields of science: indeed, it is hard to find a research area that does not rely on computer facilities. Under most applications, however, the software program conceals the underlying complexity of a computer system. Typically, a program is developed with a narrow set of specific tasks in mind, and the number of available programs grows each day. Practitioners can combine several programs to implement the desired flow of computations. At the same time, computer-processing power is also growing, thus making more complex problems accessible to researchers.

Despite the availability of a vast array of ready-to-use software, researchers still have to spend considerable time and effort to join several tools together. For solving a small-scale problem it might be appropriate to set up the software manually, but this manual setup approach becomes impractical when tackling large-scale problems. The tool setup automation is within reach of major research groups (assuming that their expert body is sufficiently diverse), but it presents a much bigger challenge for small groups with fewer experts on their staffs. Those small groups, nevertheless, also hope to benefit from the computational power available to them.

We identify two major types of tool selection problems one has to deal with when conducting data-heavy and computationally intensive research:

- Bulk data storage organization.
- Processing algorithms.

Since most researchers don't develop software tools themselves, but rather adapt commonly available software, an additional sublayer of problems emerges:

- Software selection.
- System administration.
- Multi-stage data processing and data exchange between various tasks.
- Results access and presentation.

Those problems mentioned above are all interdependent, which makes them especially difficult to solve. The choice of algorithm influences the software selection, which in turn affects the choice of the storage organization scheme and vice versa. So, even in situations when the software can effectively mask the complexity of the computer system, strong computer skills are needed for choosing the right software for the task.

The development of GRID [1] services was a first attempt to unify and consequently simplify the use of distributed computer resources. The main focus of this technology was on the consolidation of distributed computational power and distributed job execution. However, the utilization of such distributed resources is rather complex and requires strong technical skills. Besides, GRID technology specifies only a generic interface for distributed resources, irrespective of any specific problem domain. The end-user of a GRID system has limited control over external resources, which often limits his or her ability to implement more efficient solutions to their problems. Presently there exist a

Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13 - 16, 2015