ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2018, Vol. 39, No. 5, pp. 647–654. © Pleiades Publishing, Ltd., 2018.

Workflow-Based Internet Platform for Mass Supercomputing

E. V. Biryal'tsev[1*], **M. R. Galimov**[1**], and **A. M. Elizarov**[2,3***]

(Submitted by V. V. Voevodin)

[1] *Limited Liability Company "Gradient Technologies", ul. Peterburgskaya 50, Kazan, 420011 Russia*

[2] *Kazan (Volga Region) Federal University, ul. Kremlevskaya 18, Kazan, 420008 Russia*

[3] *Kazan Branch of the Interdepartmental Supercomputer Center of the Russian Academy of Sciences (RAS)—Branch of the Federal Scientific Center "Scientific Research Institute of System Studies" of RAS, ul. Lobachevskogo 2/31, Kazan, 420008 Russia*

Received October 23, 2017

Abstract—An experience of designing integrated hardware and software solutions for high-performance computing in solving modern geophysical problems on the basis of full-wave inversion is described. Problems of designing mass high-performance software systems intended for extensive use in industry are discussed.

DOI: 10.1134/S1995080218050050

Keywords and phrases: *Graphics general-purpose computing, numerical simulation, distributed computing, fullwave inversion, GPU cluster, geophysical surveys, high-performance computing platform.*

1. INTRODUCTION

At present, the development of supercomputer calculations has reached yet another qualitative milestone: the performance of high-performance clusters has cleared the bar of tens petaFLOPS, and the scientific community discusses future exaFLOPS computations [1]. The improvement of the competitiveness of enterprises and economics as a whole is largely determined by extensive mass use of supercomputer technologies [2]. In this situation, the diffusion of supercomputer systems in industry becomes more and more urgent, the more so that, at present, such systems are extensively used only by big corporations and in strategic branches.

As is known, in the 2000s, the cost and complexity of high-performance computing systems have sharply dropped. This was largely due to the creation (in 2008) and further development of the technology of graphic accelerators (GPGPU). At the end of 2017, in the World Top 500 list [3] of most powerful supercomputers (see Table 1), 85 supercomputers were equipped with GPGPU, including the Piz Daint supercomputer occupying the fired line in the rating (most powerful supercomputer in Europe), and the number of such systems increases every year.

During the seven years passed since the moment when the GPGPU technology was created, this technology has become so mature that it is now not only employed in unique machines designed for solving scientific and strategic problems but also widely used in applied computations. One of such examples is the high-performance computational system created with the participation of the authors and described below.

[*] E-mail: igenbir@yandex.ru

[**] E-mail: glmvmrt@gmail.com

[***] E-mail: amelizarov@gmail.com

Table 1. Top 500 Accelerator/CP Family System Share (November 2017)

Accelerator/CP Family	Count	System Share (%)	$R_{\max}$ (GTlops)	Rpeak (GTlops)	Cores
Nvidia Pascal	51	10.2	80 932 886	137 253 814	2 226 216
Nvidia Kepler	29	5.8	55 359 236	88 064 550	1 514 430
Intel Xeon Phi	10	2	45 466 195	71 843 547	3 918 026
PEZY-SC	5	1	22 591 662	33 063 393	23 524 704
Nvidia Fermi	4	0.8	5 510 600	10 728 261	438 916
Hybrid	2	0.4	4 144 240	7 112 717	374 764
Nvidia Volta	1	0.2	1 070 000	1 819 752	22 440

2. HIGH-PERFORMANCE COMPUTATIONS IN SEISMIC PROSPECTING

At present, oil-and-gas industry is characterized by actively seeking and adopting new technologies, which is explained, in particular, by the constant and significant increase in the cost of prospecting and production of hydrocarbon raw materials. In this situation, new, efficient and low-cost, methods of geological prospecting are searched for, which would make it possible to increase the quality of geological prediction and enhance the economic feasibility of technical decisions.

The full-wave inversion technologies [4] are promising modeling methods in geology, which make it possible to reconstruct properties of a medium in any geological conditions, unlike technologies based on the standard method of modern seismic testing, namely, the CDP (general depth point) method, which is effective when the geological medium under examination has plane-parallel structure. A special case of full-wave inversion is the technology of a passive low-frequency seismic survey of oil and gas developed in Russia [5], which analyzes the changes of characteristics of natural microseisms over oil and gas deposits. This technology is based on calculating the propagation of seismic waves in composite geological media by using grid models of dimensions up to 109 (at the present moment). The industrial application of this technology requires performing computations for numerical models of this dimension at a speed comparable with the pace of production (tens of hours) and in an economically optimal way. To accomplish this task, experts of the Kazan geophysical Limited Liability Company "Gradient" jointly with the Kazan Branch of the Federal Scientific Center "Scientific Research Institute of System Studies" of Russian Academy of Sciences and the research-and-production Limited Liability Company "Gradient Technologies" (Kazan) have performed a study, which began in 2005, aimed at the search of the most efficient soft- and hardware means for performing mass mathematical computations in geological modeling. Various hard and software platforms have been considered and special-purpose software for computations has been designed, including software using GPU clusters.

3. A SOFT- AND HARDWARE SYSTEM FOR HIGH-PERFORMANCE COMPUTING

The result of the study was the creation of a soft and hardware system built on the basis of AMD graphics accelerators. In the last (of September 28, 2015), 23rd, edition of the Top 50 list of Russian supercomputers [6], this computer system (Fig. 1), which is intended for performing geophysical tasks, placed 38th, showing performance of 35.61 teraFLOPS according to the Linpack benchmark, while built on only six nodes with GPUs. Moreover, this system is characterized in this rating by maximum performance per computational node, and the presence of only six computational nodes with 24 graphics cards witnesses that this is a compact economical solution. On the one hand, such computers are quite affordable in both price and complexity of maintenance not only for big but also for medium-sized companies. On the other hand, they have sufficient computational capacity for solving applied problems with the use of numerical models of size up to one billion cells.

The experience of designing and implementation of this system has shown that, at present, the created hardware facilities are ready for industry adoption. The direct process of assembling computational units, installing and tuning the base software involved no special difficulties. Moreover, at present, at the Russian IT-market, there are manufacturers (T-Platforms, Meijin, IntellectDigital, ARBYTE SC,

Fig. 1. 38th-placed Russian Top 50 geophysical supercomputer (September 2015).

STSS) which are ready to assembly, on order, a cluster of certain computational configuration at a price quite affordable for industrial sector. It is also possible to lease the necessary hardware, including that with GPU, in cloud systems (such as Amazon). Thus, the task of creating a technical base of mass supercomputing for industrial use, which was set in [2], can be considered accomplished to a large extent.

It appears that, in the coming decade, the mass supercomputing architecture will have the form of a relatively small (from several to several hundred) number of heterogeneous CPU/GPU nodes, in which the access of the GPU to the memory will be implemented via PCIe or NVLink, connected by a high-speed *Infiniband* network or by an *OmniPath* network, being presently developed by Intel. Increasing the availability of the SSD technology of standard SAS 12000 Mbod, which is expected in the nearest future, will make it possible to also include, according to the DAS architecture, a high-speed local storage consisting of 8–16 storage units with capacity of several ten terabytes each in every node. Thus, it is very likely that a mass supercomputer will be, according to Jim Gray's definition [7], a set of homogeneous easy to replace and augment "computing bricks" with total computing power of 0.1–10 PFLOPS and storage capacity of 0.1–10 PByte.

However, mass supercomputing, having surmounted the barrier of high price and hardware complexity, faces the problem of the complexity of hardware development. This barrier is presently being recognized by world IT-leaders; in particular, Intel has the partner *Advanced Computing Program* for supporting software development for mass supercomputing. The problem of the complexity of hardware development for computing clusters of supercomputers has two substantial aspects, the heterogeneity of the architecture of the computational platform and the distributedness of computation.

To minimize the influence of the architecture heterogeneity of clusters with GPU, methods for unifying programming the classical multicore central processors and systems with GPU mass parallelism are being developed. In this direction, most promising is the OpenCL environment, which is supported by both manufacturers of central processors and those of GPGPU. A fairly popular tool used for implementing internode interaction is the MPI protocol. The MPI + OpenCL bunch of protocols is recognized to be promising for low-level supercomputing software up to the exaFLOPS scale (see [2]). Thus, the problem of computational platform heterogeneity is being attacked fairly extensively.

The distributedness of computation, which is a priori inherent in any supercomputing project, is another important difficulty involved in software development. Traditionally, supercomputers were used to solve scientific and strategic problems at the limit of the existing technical capabilities, and main attention was given to the efficiency of numerical modeling proper. However, from the point of view of the well-known "Model-View-Controller" software model, the computational part of a numerical modeling system is only one of the three components of a program complex; the development of tools for data visualization and control in numerical models was given insufficient attention, and as a result, these components have been developed insufficiently. When supercomputer technologies were implemented for strategic branches of industry, software development was organized so as to employ big companies or research institutes possessing highly qualified personnel. This has made it possible to overcome the difficulties involved in software development for modeling both systems proper and subsystems for interaction with user and control of data.

4. THE PROPOSED APPROACH

4.1. Data Architecture

It is impossible to develop unified application software for mass numerical modeling in diverse industries. The application of supercomputer technologies in light, food, construction, and other nonstrategic industries is based on many variations of multiscale multiphysical models, the whole gamma of which cannot be developed centrally. For this reason, in these industries, software development in implemented by many small innovation companies, which cannot afford employing expensive teams of highly qualified experts. Of great importance are also the cost and terms of software development and modernization. The efficient use of supercomputer technologies in the industries mentioned above requires developing a program platform sharply reducing the complexity of designing supercomputer applications to the level of programming on personal computers. Such a platform must support not only the modeling process proper but also end-to-end processing of model data (models and modeling results).

Certainly, there exist approaches to and software for the organization of storage and visualization of large amounts of data. However, these approaches are based on specific models of data. Visualization of complex 3D scenes is extensively used in computer games, engineering problems, and modern cinematography. As a rule, software for implementing these tasks is based on the polygonal model. At present, the storage of large amounts of data has been best developed for large unstructured text/binary files or sets of such files in systems for support of social networks, electronic mail, etc.

Taking into account the large planned volume of data to be processed, we propose constructing a program platform for mass supercomputing on the basis of a unified data model in order to minimize the space occupied by data and transformations and reduce the complexity of learning. Such a structure must be suboptimal; in the case under consideration, it must be applicable, without a substantial deterioration of performance and other critical properties, to subsystems of mathematical modeling, visualization, and storage control. Below we consider models of data whose feasibility and efficiency in modeling, visualization, and storage for supercomputing in geological prospecting has been confirmed in practice.

At present, numerical modeling of physical processes is largely based on grid methods. An object to be modeled is represented by a spatial set of fixed points or cells (a grid) on which flows of physical parameters (values) are computed. This approach has been well developed methodologically and is extensively applied in modeling of aerodynamic, strength, and hydrodynamic processes, including those encountered in underground fluid mechanics and thermal and physico-chemical computations using the finite element method and its variations. The practice of solving scientific and strategic problems at the limit of technical capabilities on supercomputers required optimizing the volume of computational grids. This has led to the tendency of designing complex irregular grid constructions by using unique grid construction algorithms, which required analyzing conditionality and constructing preconditioners. Mass supercomputing requires simplifying the construction of computational grids, even if at the expense of reducing the optimality of data volume and computational time. There exists an approach to constructing universal grids with local refinement based on octrees (see, e.g., [8]).

The industrial use of numerical models in multiuser mode requires dynamically balancing the load. In solving a unique particular problem, it is possible to plan an optimal distribution of subdomains among the cluster nodes in advance, but when a computing cluster is used asynchronously by several users launching and withdrawing tasks requiring various resources at arbitrary moments of time, the number of nodes available for solving a problem must dynamically vary. Such algorithms are fairly obvious; in particular, one of solutions consists in preliminarily segmenting the computational domain into the maximum possible number of subdomains and dynamically controlling the number of adjacent subdomains processed by each node. The universal model of grids based on an octree can be applied to parallel computations [8].

The results of modeling in grid methods do not agree with the polygonal model, which is most widely used for visualization in computer graphics; as rule, a researcher is interested in the distribution of fields being modeled inside the model volume. A natural graphical representation of such objects is a voxel model, in which each voxel, that is, the elementary volume of the model space, is assigned a set of color and transparency parameters. In the case under consideration, an object already exists in the form of a grid model, and to transform this representation into a visual one, it suffices to relate the modeled parameters to color and transparency scales. Voxel graphics is applied to represent tomographic examinations, in geophysics, and to design realistic computer games. A voxel representation of objects requires much more resources in comparison with a polygonal representation of comparable precision; for this reason, in coding voxel graphics, data compression based on octrees is applied [9].

Another significant difficulty arising in visualizing results of numerical modeling is the large amount of information. In addition to the obvious problem that a model of large size computed by many nodes may be too big for a graphics card or a SLI-bunch of cards, a bottleneck is also the transmission of information from modeling nodes to the visualization node, especially when the processes to be visualized are dynamic. The visualization of numerical modeling data must be distributed. Rendering and capture must be performed directly at modeling nodes, without transmitting data to separate working places or render-farms. Thus, the data architecture of a mass supercomputing platform for numerical modeling must be based on an end-to-end data model meeting the objectives of data modeling, visualization, and manipulation, including storing and accessing data and balancing computations. A suboptimal general model is 2-trees, including quad trees for static 2D models, octrees for static 3D models, and dynamic 2D and 16-trees for dynamic 3D models. Trees of higher multiplicity can also be used to control data in spaces of versions, access levels, and other potentially possible dimensions without changing the main processing algorithms.

4.2. Software Architecture

One of the most popular trends is the hypersegmentation—or even personalization—of markets. These trends also influence software architecture which implies the ability of the software to be modified for a specific user. Formed in the era of industrial economy and focused on large-scale sample problems, classic software architecture is not suitable for personalization tasks.

Current application architecture based is hierarchic as a system-subsystem-automated workplace-function and it is focused on fixed official duties and fixed business processes. In the context of dominating object-oriented paradigm, the internal programming architecture as a system of classes and relevant methods is also focused on “fixed” condition of software products. Modification of even a small part of functionality can cause total reconfiguration of software architecture or violation of its integrity.

Development of hardware and system software tools for building an applied system is also a well-known problem. Backward compatibility of new tools with the old ones has certain limits. As a result, long-running systems operate on outdated hardware and software base.

The problem of increased adaptability and portability of applied software is one of the hot topics in IT. One can highlight several trends that are relevant to the matter under consideration.

Microservice architecture. Increased adaptability of certain functions of applied software is achieved by allocating of these functions to independent software components. Instead of implementing a function as a class method called in a single address area, functions in the microservice architecture are implemented as an external standalone program. Thus, it can be modified separately from all other software, and rewrite in another programming language using other libraries which is especially important for modification made by an executor, other than the one developing this microservice.

Lambda architecture. Another trend of increasing the adaptability of individual functions is to represent a special function as a universal one (lambda function) with a variable number of arguments. The first argument is the text of the function itself, performed in the interpreted language (script). Lambda architecture allows not only to quickly modify functions, but to modify them for each user directly in the process of execution, to personalize the functionality for each user project avoiding software adjustment. The script can be executed in a simple object-oriented language (Matlab, R,

Table 2. Comparing monolithic and microservices architectures (https://www.redbooks.ibm.com/redbooks/pdfs/sg248275.pdf)

Category	Monolithic architecture	Microservices architecture
Code	A single code base for the entire application	Multiple code bases. Each microservice has its own code base
Understandability	Often confusing and hard to maintain	Much better readability and much easier to maintain
Deployment	Complex deployments with maintenance windows and scheduled downtimes	Simple deployment as each microservice can be deployed individually, with minimal if not zero downtime
Language	Typically entirely developed in one programming language	Each microservice can be developed in a different programming language
Scaling	Requires you to scale the entire application even though bottlenecks are localized	Enables you to scale bottle-necked services without scaling the entire application without scaling the entire application

Python, Julia) available to a well-qualified expert in any domain which excludes the programmer from the process of modification of the business function.

DAG architecture. Software-based business processes performed by users consist of a sequence of individual business functions. The sequence of functions is an oriented graph (Direct Acyclic Graph), similar to activity-based network. Performed functions are nodes of the graph, while the arcs refer to data being transferred. In the classic software architecture, the distribution of functions to sites and the sequence of access to functions are threaded into the source code, which requires programmer's support in the modification of the business process. DAG architecture allows an expert in a certain domain to create a project-specific DAG from micromodules and provide it to the end users avoiding programmer. Created DAG can be saved and used as a prototype or as a typical process.

The cost of such flexibility is increased CPU time spent on software "— sometimes it is a tenfold increase compared to object-oriented architecture. However, the power of modern processors can reduce this effect to zero in terms of time and cost. When these approaches are combined, the programmer is excluded from modification of the applied part, subject-oriented algorithms are also replaced by experts in the domain, as well as parameters and input data, which enables a tenfold reduction of costs and time of modification.

4.3. Access Model

An important trend in business architecture is a business model of an Internet platform which is widely distributed in the last 5–7 years. The platform model is well known in economic theory [10]. A classic example of the platform is a city market (recently hypermarket) or a stock exchange. The main task of the platform is to minimize transaction costs related to searching a suitable product or a partner. The platform can also reduce a number of other transaction costs, i.e. produce quality control of goods, minimize the time of contract execution and guarantee fulfillment of contracts, and provide a number of specific services. A classic example of the platform is communication services (mail, telephone, Internet-based communication).

Development of Internet technologies in the last 10 years triggered the advancement of domain-specific Internet-platforms. Sale of goods is the most obvious way to apply the platform concept on the Internet. Trade platforms like eBay and Ali-Express are the giants of the world trade with capitalization of hundreds of millions of dollars. Online platform for selling software like AppStore, Uber, AirBnB are well known in more specific domains. The rapid growth of interest to Internet platforms is well traced by Scopus publications statistic for last years (see Fig. 2).

Internet platforms have proven to work in multi-side markets, which involve collaboration of several participants performing different roles. For example, GitHub, joint open software development program

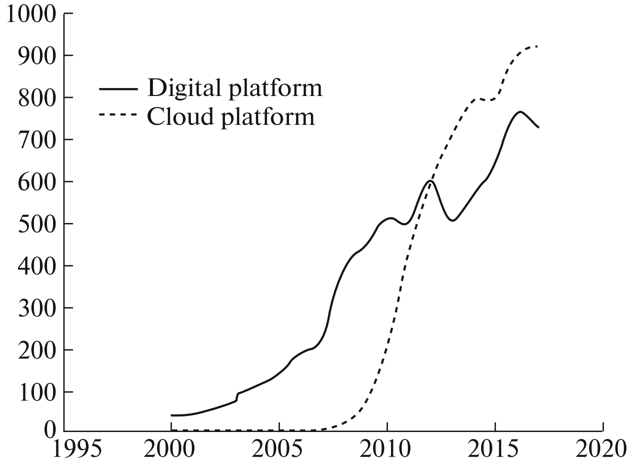

Fig. 2. Scopus publications activity.

is successfully coordinating hundreds of ongoing projects run by thousands of programmers. It is interesting to note that the capitalization of this free platform coordinating the development of open software amounted to 3 billion dollars in 2017.

In 2015, Gartner agency released an analytical report following the analysis of IT trends forecasting that in the near future we will see the emergence of a market of algorithms where the subject of transactions will be not finished software products but their components in the form of micro-modules that perform separate functions and DAG implementing typical business processes. Internet platforms dramatically reduce transaction costs, especially in multi-sided markets, making it commercially feasible to fill specific market niches.

5. CONCLUSION

One of the main consumers of the newest information-communication technologies is traditionally the oil-and-gas industry. Since the early 2010s, in oil-and gas seismic prospecting, the technology of full-wave inversion is being developed, which works in geological media of any complexity and gradually replaces he classical CDP method of seismic testing. Full-wave inversion is based on numerical methods for solving inverse problems and oriented to high-performance processing. Unlike the traditional systems for CDP processing, the technological base of full-wave inversion is at the initial stage of development. The world leaders of information market, together with the leading geophysical companies, only seek forms of organization of mass application software based on numerical modeling [11]). In Russia, integrated application software based on supercomputer numerical modeling and simultaneously including the other components of the complete cycle of data processing, such as visualization and control of data and the process of numerical computation, is being developed, too [8, 12]. The most important factor for the successful application of mass supercomputing is a software platform which is as easy to master as possible and based on a model of data common for the entire data processing cycle and suboptimal for computation, data control, and visualization. Universalization of data structures will make it possible to reduce transformations of many-terabyte data arrays in the system and, most importantly, maximally facilitate learning the software platform in applications and speed up the development and modification of applied software.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 18-07-00964, and was funded by the state assignment to the Joint Supercomputer Center of the Russian Academy of Sciences for scientific research, project no 0065-2018-00015.

REFERENCES

1. G. da Costa et al., “Exascale machines require new programming paradigms and runtimes,” J. Supercomput. Front. Innov. **2** (2), 6–27 (2015). doi 10.14529/jsfi150201
2. V. B. Betelin, E. P. Velikhov, and A. G. Kushnirenko, “Mass supercomputer technologies—the basis of competitiveness of the national economy in 21st century,” Inform. Tekhnol. Vychisl. Sist., No. 2, 3–10 (2007). http://www.jitcs.ru/index.php?option=com_content&view=article&id=178.
3. http://www.top500.org/.
4. P. R. Haffinger, *Seismic Broadband Full Waveform Inversion by Shot/Receiver Refocusing* (Delft Univ. Technol., Delft, 2013). http://repository.tudelft.nl/view/ir/uuid%3Ad2d8d264–5037–4573–8418-a079afa8d1e7.
5. N. Ya. Shabalin, V. A. Ryzhov, and E. V. Biryal'tsev, “Passive low-frequency seismic survey—myths and reality,” Prib. Sist. Razved. Geofiz., No. 2, 46–53 (2013).
6. http://top50.supercomputers.ru/?page=rating.
7. The Fourth Paradigm: Data-Intensive Scientific Discovery. http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf.
8. Yu. V. Vasilevskii, I. N. Kon'shin, G. V. Kopytov, and K. M. Terekhov, *INMOST-Software Platform and Graphical Environment for Developing Parallel Numerical Models on Grids of General Form* (Izd. Mosk. Univ., Moscow, 2012) [in Russian].
9. T. Karras, S. Laine Samuli, and G. J. Ward, Efficient Sparse Voxel Octrees-Analysis, Extensions, and Implementation (1984). https://mediatech.aalto.fi/samuli/publications/laine2010tr1_paper.pdf.
10. S. Choudary, M. Van Alstyne, and G. Parker, *Platform Revolution: How Networked Markets are Transforming the Economy—and How to Make Them Work for You* (W. W. Norton, New York, 2016).
11. Global Technology Leader in Visualization and Visual Compute. http://hue.no/sites/default/files/Hue_Brochure_US_web.pdf.
12. E. V. Biryal'tsev, P. B. Bogdanov, M. R. Galimov, D. E. Demidov, and A. M. Elizarov, in *Proceedings of the National Supercomputer Forum, Pereslavl-Zalesskii, Russia, 2015.* http://2015.nscf.ru/TesisAll/8_Integraciya_visokoyrovnevix-_resyrsov/07_476_BiryaltsevEV.pdf.