On the Rostrum of the RAS Presidium

The Prospects of Development in Russia of High-Performance Computing and Predictive Modeling in Modern Technologies

B. N. Chetverushkin^{a,*} and M. V. Yakobovskiy^{a,**,#}

^a*Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, Russia*

^{*}*e-mail: chetver@imamod.ru*

^{**}*e-mail: lira@imamod.ru*

Received May 1, 2021; revised June 2, 2021; accepted August 16, 2021

Abstract—This article is dedicated to the development and application of supercomputers and supercomputer technologies, which constitute a major factor in scientific, technical, and social advance, for solving problems that require large computing capacities. Trends in the use of supercomputers are considered, and the level of equipping advanced countries with supercomputer resources is discussed. Examples of tasks the solution of which is impossible without using computing capacities of tens of petaFLOPS are given. This article was prepared on the basis of a report presented at the meeting of the Presidium of the Russian Academy of Sciences on February 16, 2021.

Keywords: mathematical modeling, supercomputer technology, pilot tasks, predictive modeling, shared center, big data, national security.

DOI: 10.1134/S1019331621060095

The intense growth in the performance of computing systems is confirmed by the fact that mathematical modeling on computing systems of high and ultrahigh performance, including processing big data, is a major factor in scientific, technical, and social advance. According to the top-500 list [1] for November 2020, the Japanese supercomputer Fugaku has a peak performance of 537 PFLOPS,¹ which is almost three times higher than the performance of the leader of the November 2019 list.

Using supercomputer technologies, it is possible not only to obtain new knowledge but also to create new equipment and technologies in a much faster and less costly way. RAS meetings held with the participation of representatives of academic science and designers general of the military—industrial complex confirm the urgent need to use computer modeling methods in the development of new technology. This is especially true in the context of the ban on the transfer of materials and technologies to Russia by Western countries. Another area of applying supercomputer technologies is the processing of available large amounts of information, which has a significant

impact on making informed decisions in various spheres of public administration. Supercomputer technologies are an important factor in ensuring Russia's national security. Table 1 shows some high-performance computing systems in the world and in Russia [1].

The general service meetings of the core group held at the Institute of Applied Mathematics under the leadership of Academician M.V. Keldysh used to debate the problem of how much machine time to allocate to one group or another or to one department or another: “Imagine we had had 10–15 MFLOPS (note that at that time our main megaFLOPS computer was BESM-6)—we would have solved all the problems.” Keldysh, with his usual kind ironic smile, would answer, “If you had had 15 MFLOPS, in six months you would come and say: give us 100 MFLOPS. With 100 MFLOPS, in six months you would want GFLOPS.” This tendency is characteristic of the present day as well.

Note some rows of Table 1. Position 7 presents information about the computing system of the supercomputer center of the German city of Jülich, which has a peak performance of 70 PFLOPS. At position 10 of the list is the computing system of Saudi Arabia, which has a peak performance of 55 PFLOPS. The presence of these systems shows a high demand for supercomputer technologies in all areas, not only basic but also purely commercial. Plans have already been announced to reach the 1 EFLOPS (10^{18} FLOPS)

[#] RAS Academician Boris Nikolaevich Chetverushkin is Director for Science of the Keldysh Institute of Applied Mathematics, RAS (KIAM RAS). RAS Corresponding Member Mikhail Vladimirovich Yakobovskiy is Deputy Director for Science of the same institute.

¹ One petaFLOPS corresponds to 10^{15} floating point operations per second.

Table 1. List of high-performance computing systems

Position in the list	Name	Country, year of recording in the list	Peak performance, PFLOPS	Power consumption, MW
1	Fugaku	Japan, 2020	537	29.9
2	Summit	United States, 2018	201	10.1
3	Sierra	United States, 2018	125	7.4
4	Sunway TaihuLight	China, 2016	125	15.4
7	JUWELS Booster Module	Germany, 2020	71	1.8
8	HPC5	Italy, 2020	52	2.3
10	Dammam-7	Saudi Arabia, 2020	55	n/a
40	Christofari	Russia, 2020	6.6	8.8
156	Lomonosov 2	Russia, 2018	4.9	n/a
500	Internet company T FF1	China, 2018	2.4	n/a

milestone by 2023. The installation of such systems is planned by the United States, the European Union, and China. NASA plans to achieve 30-EFLOPS performance by 2030.

The largest computing system in Russia, Christofari, installed in Sberbank, with its performance of about 9 PFLOPS, stands in contrast to this background. The most powerful Russian system is only in 40th place in the world ranking. The only other domestic system included in the list occupies 156th place, its performance being 4.9 PFLOPS. Note that the bottom 150 positions (performance of about 1.5 to 2 PFLOPS) are occupied by systems of many countries, including China (59 systems), the United States (39 systems), the Netherlands (15 systems), Ireland (14 systems), Japan (five systems), Britain (four systems), and Singapore (four systems). There are systems of this class in Canada, Norway, France, Taiwan, the Czech Republic, Brazil, and Hong Kong. Today they serve as a working tool used to solve current problems, including the following:

- 26 systems, in the research and educational sectors;
- 11 systems, in the public administration sector;
- 108 systems, in the industrial sector;
- 13 systems, in the academic sector.

Well-deserved compliments should be addressed to the management of Sberbank. In fact, those individuals sensed in time the upcoming transformation of the banking sector, inextricably linked with the intensive use of high-performance systems.

Back to Germany. We have been cooperating very closely for many years; for several years in a row, we held symposia together with the leadership of German computing centers. A whole line of supercomputer

centers has been created in Germany: in Jülich, Stuttgart, and other cities, each center having its own specialization. The supercomputer center in Jülich serves as the headquarters. Numerical methods are being worked out there, and pilot tasks are being solved. Dresden specializes in modeling social projects and solving big data problems; Hamburg, in high energy physics; Munich, in energy problems; and Stuttgart, in the aerospace and automotive industries and urban ecology.

For example, Mercedes-Benz simulates various components and technologies and solves problems related to laser welding, combustion, etc. The Stuttgart center is equipped with effective means of 3D imaging of the results of supercomputer modeling, which provide an opportunity for visual perception and comprehensive study of multidimensional processes. You enter the combustion chamber and see streams of hot gases of different concentrations flowing around. You can see with your own eyes the peculiarities of the distribution of the key parameters and determine the directions for optimizing the structures to ensure the required operating modes, for example, to reduce the emission of harmful gases.

Indicatively, three or four years ago, the characteristic performance of the main supercomputer centers in Germany was at the level of 5–10 PFLOPS, while now it has already reached the level of 25–50 PFLOPS. In Germany, about ten such systems specialize in research for various industries.

Supercomputer systems are also of considerable interest from the standpoint of expanding the possibilities of carrying out not only applied but also basic research. Let us consider a few examples that illustrate our potentialities and show what we lose if we do not

Fig. 1. Results of modeling the dynamics of the absorption of galaxy matter by a black hole [2].

have computing systems available to foreign specialists.

$$\begin{aligned} \frac{d\rho}{dt} + \operatorname{div} \rho \mathbf{u} &= 0, \\ \frac{d\rho \mathbf{u}}{dt} + \operatorname{div} [\rho \mathbf{u} \times \mathbf{u} + B_k B_p] + \nabla \left(p + \frac{\mathbf{B}^2}{8\pi} \right) &= \operatorname{div} P_{NS}, \\ \frac{d\rho E}{dt} + \operatorname{div} \left[\left(E + p + \frac{\mathbf{B}^2}{8\pi} \right) \mathbf{u} \right] &= \operatorname{div} \mathbf{q} + \operatorname{div} P_{NS} \mathbf{u}, \quad (1) \\ \frac{d\mathbf{B}}{dt} &= \operatorname{rot} \mathbf{u} \times \mathbf{B} + \operatorname{rot} v_m \operatorname{rot} \mathbf{B}, \\ \operatorname{div} \mathbf{B} &= 0, \\ \Delta \Phi &= 4\pi G \rho. \end{aligned}$$

Relations (1) describe the system of equations of magnetic gas dynamics, supplemented by account for the gravitational potential. Such an addition makes its solution an order of magnitude more difficult since it requires joint modeling of different-scale processes. Figure 1 shows our calculation of the absorption of galaxy matter by a black hole [2]. Here we see for the first time how a narrowly directed cosmic plasma jet is formed. Such jets have been observed by astrophysicists in a number of stellar systems (a typical example is shown in Fig. 2). Without going into the details of the computational experiment, we should note that the calculation was carried out by domestic specialists, but on a foreign computer (this opportunity was provided by the Supercomputer Center of Hamburg). The right side of Fig. 1 corresponds to a 3D calculation performed on 4 bln spatial calculation points. If the calculation is performed on a less detailed grid (by 500 mln), then the result shown on the left side of Fig. 1 will be obtained, in which the effect of the appearance of a jet is not observed. A vortex does not

form, and the magnetic fields that push out the flow of particles do not appear. Without high-performance computers, we cannot solve many problems of basic science with the help of a computational experiments or detect many significant effects.

This concerns not only problems of astrophysics. It is widely known how important it is to consider the different scales of turbulence together. At any transi-

Fig. 2. Jet of the galaxy M87 [3].

Fig. 3. Dynamics of air flow around a helicopter rotor.

tion from a larger to a smaller scale of turbulence, the grid in each direction should be increased by about five times. Solving combustion problems also requires large computing capacities. It is necessary to take into account both gas dynamics and chemical kinetics—processes with completely different characteristic flow times. To calculate a jet engine with 500 mln nodes only with energy, without ecology, it takes three to four days of work on a PFLOPS computer. If we consider ecology (note that now everyone is fighting for clean exhaust), it is necessary to take into account a long chain of chemical reactions, which greatly increases the complexity—tens of petaFLOPS are required. This means that without high-performance computers, we lose a lot both in basic science and in improving industrial technologies.

Let us give an example of calculations performed by our specialists [4, 5] for the dynamics of interaction with the air of a working helicopter rotor blade. Despite the development of special measures to reduce the grid around each blade, the predictive modeling of vortex separation; noise generation mechanisms; and determination of the lift force, which largely depends on the blade profile, detailed grids are required containing about hundreds of millions of spatial points. If you increase the detail of the grid in each direction by three times, you will have to increase the time of the 3D calculation by at least 27 times. In addition, recall the inevitable decrease in the time step owing to the reduction in the cell size of the computational grid. In total, the calculation time increases by 81 times or more. The calculations presented (Fig. 3) were performed at the limit of the capabilities of a computing system with a performance of 0.2 PFLOPS. Thanks to the methods developed, we can effectively use supercomputers with a performance of tens of petaFLOPS, but our specialists do not have sufficient access to such computing systems.

Note the research carried out at NASA. The specialists of this agency solve problems using billions of points. Characteristic computing power costs are mil-

lions of core hours² for each option. True, there is a certain nuance here. In solving optimization problems, it becomes necessary to perform many independent series of calculations, which greatly facilitates the problem of the efficient use of large computing capacities. It becomes possible to limit each calculation option to a small number (tens of thousands) of processor cores, simultaneously launching many independent calculation options, each with its own values of the parameters under optimization. There is an opportunity to use a larger number of cores (hundreds of thousands) for each calculation, but this reduces the efficiency of using the processors. In some cases, this is justified, but when solving optimization problems, both from the point of view of economy and from the point of view of reducing the total computation time, it is more profitable to run many options simultaneously—each on a relatively small number of processor cores. This method is not suitable for solving with high accuracy large, tightly coupled problems, such as those considered in the first two examples. They need a large number of processor cores and accelerators, simultaneously working on one calculation option.

Interesting work, including from the point of view of ecology, is carried out at the RAS Nuclear Safety Institute (IBRAE) and the RAS Institute of Numerical Mathematics (INM) on modeling the transfer of radionuclides in storage facilities of radioactive materials. Each calculation uses tens of millions of calculation points. The computing capacity available in Russia makes it possible to make a forecast for 10 000 years. However, foreign colleagues, studying the storage facility of the Forsmark Nuclear Power Plant (Sweden), perform calculations using billions of points and make a forecast for a million years. Thus, in ecology, we also state a significant lag due to the lack of powerful computing systems.

Noteworthy are fashionable trends such as artificial intelligence and big data. Let us start with digital nutritionology. The corresponding work began on the initiative of Academician V.A. Tutel'yan together with the Keldysh Institute of Applied Mathematics, RAS, and the Federal Research Center of Nutrition, Biotechnology, and Food Safety [6–8]. It is aimed at the formation of a balanced diet with account for various factors, including age, gender, income level, ethnic characteristics, physical activity, chronic diseases, and dietary diversity. The benefits of oatmeal are undoubted, but you cannot eat oatmeal alone. With an increase in the number of factors taken into account by the mathematical model by two times, the number of necessary calculations increases tens of times. Our preliminary calculations show that, if we consider the population of Russia as a whole, then to determine the composition and plan the production of the necessary food products, computing resources of about a day are

² A core hour refers to the number of computational operations performed by one processor core per one hour of calculation.

required on a system with a performance of 10 PFLOPS. There is a need to carry out calculations for the construction and analysis of such forecasts. They are necessary for medicine and food logistics. The developed methodology can be applied to planning other mass productions as well.

Meanwhile, the Japanese Fugaku system will largely be used to simulate the processes associated with the coronavirus pandemic. The transnational oil and gas company British Petroleum is providing a 16.3 PFLOPS supercomputer in Houston for modeling COVID-19. As part of the research carried out on it, various aspects of modeling are considered: the penetration of the virus through a mask, multifactorial modeling of the impact of certain measures on population groups, etc. The amount of computation required significantly and nonlinearly depends on the size of the region: a city with a population of 100 000 people and a city with a population of one million are different pairs of shoes. With an increase in the size of the population group considered by ten times, the volume of calculations changes not by ten times but by much more. Multifactorial modeling is very demanding on supercomputer resources.

Another urgent task is to monitor public sentiments based on the analysis of Internet messages. We are now doing such work together with our colleagues from Snezhinsk. The methods created make it possible to identify, based on Internet messages, COVID alarmists, COVID dissidents, those who have become extremely neurotic by the situation with the pandemic, and those who understand the danger of going out but have to go out for economic reasons. All this is extremely important for assessing the effectiveness and consequences of management decisions. Obviously, the more factors one can take into account and the more detailed the description of the population groups, the more reliable the forecast will be, but its implementation requires greater computing capacities than those currently available to domestic specialists.

An important area of application of supercomputers is strategic planning of traffic flows, ensuring the connectivity of the territory of the Russian Federation. We are working on planning the optimal use of the aircraft fleet and a network of small transport hubs. It is not necessary to fly through Moscow. Currently, we are using 0.15 PFLOPS systems, but they do not allow us to build models commensurate with the scale of the country. If we want to consider large territories, especially the Far North and Siberia, we should take into account not only the aircraft fleet but also road, rail, and river transport. Again, in this case, the amount of computation increases dramatically, and large computing capacities are required for planning and forecasting [9].

Disaster analysis, risk assessment, and countering unforeseen situations are becoming an interesting area of research. Here, neural networks and machine learn-

ing are used to make quick decisions. However, machine learning works well when a lot of input data are available. To fill the training samples, calculations of rather complex virtual disaster scenarios are carried out. In particular, the IBRAE RAS considers regular and abnormal operating modes of reactors, the simulation of which also requires large computing capacities.

Another promising area of significant interest for intensive development is the Safe City project. Its implementation is associated with assessing public sentiments and countering possible anthropogenic and natural disasters. Complex accounting of many factors in real time requires significant computing resources. The Digital Hydrocarbon Field project and many other tasks that require computing capacities of tens of petaFLOPS for their solution fall into the same class of resource-intensive tasks.

Note some fundamental points. As was mentioned above, it is easier to compute several scenarios simultaneously running not one but several calculation options. Such a strategy, generally speaking, increases the total number of calculations since effective optimization algorithms, when planning the next option studied, consider the entire amount of information obtained previously. When calculating several options at the same time, it becomes necessary to plan several options simultaneously. Thus, for some options, less information is used in planning than compared to sequential calculation. Therefore, some of the calculations may be less "useful" or even redundant. Nevertheless, specialists prefer to put up with such costs since they are directly determined by the existence of a complex problem of adapting algorithms to the architecture of systems with extramassive parallelism. When a lot of computing cores are involved, they interfere with each other, like a crowd of people walking through a narrow door. Logically simple and efficient algorithms are needed, but they are difficult to find. Now the intensive search for effective methods is being conducted. In fact, this is the task of basic science and basic aspects of applied mathematics. Note that, in this area, Russian scientists are not lagging at all, which is why foreigners cooperate with us. This is our competitive advantage, but if we do not have high-performance computers, we will lose it.

This statement is not that without supercomputers nothing works and will not work. Even using the existing equipment, we manage to solve many problems through methods, through work with foreign partners, or through the use of simpler but adequate mathematical models. However, we are approaching the limit of such possibilities. We inevitably miss a number of nuances, many of which are essential. For further development and maintaining parity with foreign competitors, high-performance computers are needed.

Russia, by virtue of the logic of scientific and technological progress and its geopolitical position, should

dramatically increase the productivity of its computing park. It is necessary to bring it at least to the level of the line of German computing centers. However, in terms of our geopolitical ambitions, we are not Germany. This must be clearly understood. In Germany, there are a dozen 25–35 PFLOPS centers and some with a performance of about 50 PFLOPS, and we have only the Christophari supercomputer of Sberbank and a slightly smaller supercomputer, Lomonosov, at Moscow State University. Without developing supercomputer resources in the proper measure and at the proper pace, we are doomed to technological lag in all strategically important areas. Of course, a computational experiment is not the only tool for studying the properties of the surrounding world and developing new technologies. There is also the field experiment, there are also the brains of our physicists and chemists; still, without sufficient computing capacities, we rob ourselves and deprive ourselves of many opportunities.

We need to create a line of high-performance centers. Significant regional and university centers should be put into operation. Moreover, one of the large supercomputer centers, by analogy with the center of the German city of Jülich, should be focused on the development of fundamental methods for solving computationally complex multiscale multidisciplinary problems. Researchers of various specialties, concentrated in the Russian Academy of Sciences, are necessary for developing methods and solving pilot tasks. The supercomputer center of the Russian Academy of Sciences, perhaps not the largest, but still powerful, lagging behind the first positions of the Top-500 by no more than an order of magnitude, can play the role of a kind of experimental reactor that provides new knowledge and new high-performance computational methods and technologies. This will not only make it possible to solve complex problems successfully but also ensure the further development of domestic high-performance software products, the market of which is currently extremely far from saturation not only in Russia but also in the world. This will ensure the preservation, use, and further development of the competences of Russian scientists and specialists.

Funds for the development of computer technology and domestic hardware components must be found. The rightmost column of Table 1 indicates the electrical power consumed by a computer (not by the center but by the computer; the center as a whole requires much, at least a third, more). The power consumption is measured in megawatts, which makes it costly not only to create but also to operate supercomputer centers. Nevertheless, this is a case when it is inadmissible to cut costs. Let us recall how the nuclear industry was created in our war-torn country. Since then, we have used the security then provided and the technologies created then, spread far beyond the industry. It can be argued that then there was a different social system and a different attitude towards human rights. Yes, indeed, times have changed, but

here is another example—the pandemic. At the state level, funds have been allocated, hospitals and vaccines have been created, and support to small and medium-sized businesses has been provided. We are passing these hardships with dignity largely due to the fact that the country's leadership has allocated proper resources.

Summing up, let us note once again: the main thing today is to realize the danger. We constantly hear from representatives of power structures: “Just show that these high-performance computing systems are really necessary, what they are for, and what economic effect they will bring.” First, the loss of an adequate level of equipment with supercomputer resources leads (because of the degradation of the level of key digital technologies) to an uncontrolled growth of threats to Russia's national security. Second, the benchmarks set by technology leaders cannot be ignored. It is doubtful that the world's leading powers are wasting large sums of money, equipping supercomputing centers with advanced computing technology and making it widely available to research, university, and commercial organizations. Third, the opinion of the Russian Academy of Sciences, the leading expert organization in Russia in the field of science and technology, is unambiguous: high-performance supercomputer systems are necessary for the country's development.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. List of 500 supercomputers, November 2020. <https://top500.org/lists/top500/2020/11/>.
2. B. Chetverushkin, N. D'Ascenzo, A. Saveliev, and V. Saveliev, “A kinetic model for magnetogasdynamics,” *Math. Models. Comput. Simul.* **9** (5), 544–553 (2017).
3. Astronomy Picture of the Day 2000, July 6. <https://apod.nasa.gov/apod/ap000706.html>.
4. V. G. Bobkov, I. V. Abalakin, and T. K. Kozubskaya, “Method for prediction of aerodynamic characteristics of helicopter rotors based on edge-based schemes in in code NOISEtte,” *Comput. Res. Model.* **12** (5), 1097–1122 (2020).
5. I. V. Abalakin, V. G. Bobkov, T. K. Kozubskaya, et al., “Numerical simulation of flow around rigid rotor in Forward flight,” *Fluid Dyn.* **55** (4) 534–544 (2020).
6. M. B. Gavrikov, A. A. Kislitsyn, Yu. N. Orlov, et al., “Numerical personal threpsology: The problems and solutions,” *Keldysh Institute Preprints*, No. 25 (2020).
7. Y. N. Orlov, A. A. Kislitsyn, A. O. Kambarov, et al., “Digital Nutrition: Spectral portraits of optimal diet,” *Sci. Vis.* **12** (2), 139–150 (2020).
8. V. A. Tutel'yan and D. B. Nikityuk, *Nutritiology and Clinical Dietology* (GEOTAR-Media, Moscow, 2020) [in Russian].
9. B. N. Chetverushkin and V. A. Sudakov, “Factor modeling for innovative-active enterprises,” *Matem. Mod.* **32** (3), 115–126 (2020).

Translated by B. Alekseev