



SUPER COMPUTERS

DIGEST 2010-2012

Chairman of the Editorial Board

Vladimir VOEVODIN
Voevodin@parallel.ru

Editor

Igor LEVSHIN
Igor.levshin@supercomputers.ru

Art Director

Victoria IVASHKOVA

Proofreader

Martyn ANDREWS
mandrews@rttv.ru

CEO&Publisher

Daniel ORLOV
Publishing House SCR-Media
Daniel.orlov@supercomputers.ru

Business development officer

Andrei CHELYSHEV
Andrei.chelyshev@supercomputers.ru

117342, Moscow, ul. Butlerova 17b
www.supercomputers.ru

© Publishing House SCR-Media 2012



Dear friends!

Welcome to the English language version of the Russian magazine «Supercomputers» – the only magazine in Europe dedicated solely to the topic of HPC. We translated some of the articles which we found to be the most interesting. They are the most up to date topical reports and older features whose importance hasn't changed. Emerging technologies, development, problem solving, facilitation and how the whole world is changing! We didn't change anything editorially or update information – we will let you discover and compare how our authors with the highest level of knowledge predicted the growth and trends of

the supercomputing industry. You will see for yourselves the adjustments that have been achieved in reality, the alterations and the progress made. Russia has a long history of parallel computing. Today the reality is the integration of the practical development in HPC within each individual country with its IT culture and practical results into the united European experience. The mission of the magazine is to discuss effective communication between all HPC market players, consumers and scientific organisations. With this, maybe the world will be a better place? Enjoy!

Vladimir Voevodin

CONTENTS

HPC: Where are we?	4
Editorial interview: Dr. Hans Werner Meuer	
An HPC workhorse	10
Supercomputer for Superjet	12
Predicting the future: TOP500 2018, what it will be like?	16
Modern software and hardware solutions for aircraft engine applications	20
Swallow in the clouds	24
Editorial interview: Pr. Satoshi Matsuoka	
Cooling systems of the future	26
Water-cooled machines: another false start?	30
More work must be done in less time	34
Editorial interview: Pr. Thomas Sterling	
Cleversafe: revolution in storage	36
SciDB – new DBMS for large amounts of scientific data	41
On a motorway with two lanes	44
Editorial interview: Coordinator Jean-Yves Berthou	
Siberian Federal University (SFU)	46
Computer numerical mathematical model and computer analogy method	50
Will quantum computers take on conventional ones?	54
From «Dragonfly» to the «ZETA SCALE»	58
Editorial interview: Expert Steve Scott	
Eddy bearings and liquid balls: virtual and real	60
Hadrons and human health	65
Supercomputer in theoretical medicine	68



ISC'12 THE HPC EVENT

June 17–21, 2012, Hamburg, Germany

Join the Global Supercomputing Community

Top 10 Reasons to Attend ISC'12

- A great venue for networking
- Unmatched access to finest minds
- Information that you can apply when you return to work
- Source of inspiration for new ideas and new ventures
- Sessions packed in NEW dynamic formats as never before seen
- Join the TOP500 winners
- Your next customer could be here
- Meet with the specialists and solve your problems on the spot
- Shape your future career
- Hamburg is a fun city



Tutorials and Workshops

June 17, 2012

Tutorials

- Full and half-day tutorials
- Cover a broad range of HPC areas of interest

HPC in Asia Workshop

- HPC systems, performance, applications and research in Asia

Conference Highlights

June 18 – 21, 2012

- Energy Efficient HPC Centers
- Future Heterogeneous Architectures
- Supercomputer Architectures for Data Intensive Applications
- Alternative Processors, Architectures and Multidisciplinary Applications
- Best Practices of Large-Scale Applications across Industries
- Petascale Systems in the World and their Applications
- Application Performance: Lessons learnt from Petascale Computing
- Panel: Programming Models in the Years to come
- Critical Aspects of High Performance Networking
- The Realities and Challenges of HPC in the Cloud
- Exascale Computing: Where we are?
- Exascale Chats with Horst Simon, Yutong Lu, Mateo Valero, Mark Seager
- Networking/Interconnect within HPC-Systems
- Analyst Crossfire & Think Tank with HPC Luminaries
- Parallel File Systems
- Computational Chemistry
- HPC for Small and Medium Sized Enterprises
- Publication of the 39th TOP500 list

ISC EXHIBITION

June 18 – 20, 2012

- Supercomputing, Networking and Storage solutions
- 160 industrial, research and start-up exhibitors

CONFERENCE

June 18 – 21, 2012

Contributed Sessions

- Research Papers & Posters
- BoFs

Industry Sessions

- Exhibitor Forums
- Hot Seats

Satellite Events

June 17 – 21, 2012

- PRACE Scientific Conference
- GAUSS Symposium
- HPC Advisory Council Workshop
- CADFEM Workshop
- IDC Breakfast Briefing

The 2012 Keynotes

Monday, June 18 | 9:30 am

Conference Keynote

Tuesday, June 19 | 5:15 pm

Wednesday, June 20 | 5:15 pm

Thursday, June 21 | 12:15 pm

- **HPC as Enabler for the Virtualization of Aircraft Development**
Guus Dekkers, *CIO of EADS and Airbus*
- **Advanced Memory Technology – #1 Factor for Energy Efficient HPC**
Dr. Byungse So, *SVP, Head of Memory Product Planning & Application Engineering Team, Samsung*
- **HPC Achievement & Impact 2012**
Prof. Dr. Thomas Sterling, *Professor of Informatics & Computing, Indiana University*
- **Extreme Energy Efficiency with SuperMUC**
Prof. Dr. Arndt Bode, *Director, Leibniz Rechenzentrum & Technische Universität München*



Platinum Sponsors



Gold Sponsors



Silver Sponsors



Media Sponsors



HPC: Where are we?

Editorial interview

Published: #10 Summer-2012

Editor of our magazine, Igor Levshin, has interviewed the General Chair and founder of International Supercomputing Conference (ISC) Dr. Hans Werner Meuer.



Igor Levshin: You are often speaking about an increase in the number of people coming to ISC. What about countries – which countries are new to the conference?

Hans Meuer: ISC'12 will be our 27th conference on Supercomputing, Storage and Networking. In the last 26 years, we have had the pleasure of witnessing an annual increase in conference attendee and exhibitor participation. As for ISC'12, we are hopeful that we'll reach our goal of 2,400 participants compared to 2,200 in 2011. Last year's attendees came from 55 countries and among them was a small percentage from new countries in the Middle East, Asia and South America.

I. L.: Do you see any changes in the geography of visitors, speakers, and booth owners?

H. M.: The largest group of participants still comes from Germany, but compared

to the past, this share has declined sharply in the recent years. We also see a strong commitment from Asia, especially China which trumps with an above-average rise in the number of exhibitors. An even stronger growth is recorded by Russia, which forms ISC's fifth largest participating country after Germany, the U.S., U.K. and France. We are also proud to host five Russian exhibitors from the research and industry sectors at ISC'12. And for the very first time, we also have an exhibitor from India.

I. L.: You stress the importance of the Asian Workshop at the conference.

Are you thinking about organizing an event in Asia?

H. M.: Presently, we don't have any intention of expanding our base but who knows what the future holds in store. However, we will continue to support and promote Asia's HPC achievements at our own conference. We acknowledge Asia's supercomputing capability and know-how in sessions such as the Research Paper, Heterogeneous Architecture, Energy and HPC, Exascale Computing: Where Are We?, File Systems, New Petascale Systems. The second HPC in Asia workshop is fully dedicated

to Asia, and we are glad that India will also be represented at this year's workshop.

I. L.: Do you feel that «Big Data» is becoming an increasingly important topic nowadays? Does it deserve its dedicated events and workshops at ISC?

H. M.: Traditionally, number crunching ruled HPC since the very beginning with the Cray 1 in 1976. For 20 years, the TOP500 ranked the (number crunching) supercomputers in the world by their best Linpack performance. Very recently, with «Big Data» this has changed in science



and industry. Data crunching has soared in importance and visibility gaining a comparable interest as number crunching. A new approach for ranking (data crunching) supercomputers has been set up with the Graph500 project. Rather than FLOPS, the used yardstick is TEPS (Traversed Edges Per Second). The Graph500 project is still at the beginning with a list of just 49 distinct supercomputers in November 2011. At ISC'12, the Graph500 project and the TOP500 project with new developments will be discussed in the session «New Developments for Ranking Supercomputers». That the HPC manufacturers are beginning to look at new architectures for data crunching machines will be demonstrated in a session on «Supercomputer Architectures for Data Intensive Computing». Cray's CEO Peter Ungaro and Convey's Chief Scientist Steve Wallach will introduce their Supercomputer Architectures for Data Intensive Computing, thus demonstrating that 'Big Data' is becoming an increasingly important topic now.

I. L.: Is supercomputing based on «Light weight cores» (like ARM and others) becoming another focus of user interest? Do you expect a surge of reports here?

H. M.: My answer is «yes» and the reason for it is that many-core chips have changed high-performance computing systems tremendously. Because cache coherence is difficult to maintain across many cores, manufacturers are exploring designs that do not have any cache coherence between cores. Communications on such chips is implemented using message passing, which makes them resemble like clusters. Special hardware can be provided that supports very fast on-chip communications, reducing latency and increasing bandwidth.

I. L.: What's the current attitude toward «Cloud computing»? Will it become an area of business and technology very separate from

supercomputing or will we see a convergence in future?

H. M.: To answer this question, we would need to distinguish between top-end supercomputing and mainstream HPC, because the answers are different.

Top-end supercomputers are at least one order of magnitude faster (if not more) than mainstream systems, performing at about 1 Petaflops Linpack and up. We might see about 15-20 of these top-end systems in the next TOP500 list in June, currently starting at a price point of about \$15-\$20 M.

Because they are mainly used for big-science applications, these systems are usually running at full load. For such scenarios, we can easily calculate the \$/hour computing cost, which is similar to that of today's commercial HPC clouds. Thus, there is no major additional benefit from commercial HPC clouds for these top-end systems other than perhaps a greater flexibility. And this, by the way, holds for all HPC systems which are running under full load!

Now let's look at mainstream HPC, namely all HPC systems which are used for executing daily workloads like in automotive design, oil and gas, chip design, drug design, and so on.

As soon as these systems are not continuously fully loaded (and this is project dependent), some portion of their compute cycles are not used, and thus they are wasted. The larger this portion, the better suited are HPC Clouds! This is because in a commercial HPC Cloud, you are usually sharing the HPC resource with other users, leading to a fully loaded HPC Cloud, and the cost per cycle can be minimized.

The ideal HPC Cloud scenario for this situation is the hybrid cloud: you buy a smaller HPC system for your private cloud covering the average load, and you use the commercial cloud for any peak demand. This will result in cost savings, faster response time, and increased business flexibility – to spell out the least.

I. L.: How can interests of business leaders be leveraged with scientific and open source look and feel in Cloud Computing? Do you see any changes in attitudes and relations?

H. M.: Open source usually drives standards, interoperability, and flexibility. This is beneficial for business which need to run their applications on different clouds: their applications, data, and results are easier to move from one cloud to another. This is very useful when security, reliability or choice are the strong requirements: for example, you can run an application with different parameters on different clouds thus keeping the overall combined solution a secret; you are not vulnerable if one cloud provider goes out of business; or, with an open choice of cloud providers you have a choice of quality and price. In these and similar scenarios, open source is useful. But, it has to be mentioned that several of these benefits can also be achieved by simply offering (and using) standard APIs for easily getting in and out of a cloud.

I. L.: Did you find any particular technology or economic changes in HPC revolutionary last year? What was most exciting at the last ISC and what new developments do you expect at ISC 2012?

H. M.: There are few happenings from 2011 that are worth mentioning here. The first is that many bigger companies acquired smaller companies, for example Platform Computing, which is a leader in cluster, grid and cloud management software was purchased by IBM, Intel acquired the InfiniBand business of Qlogic, Mellanox Technologies made itself more competitive by purchasing Voltaire, a provider of scale-out data center fabrics, Dell bought over Force10's networks and Hitachi Data System acquired BlueArc, to strengthen and improve their offerings.

The second highlight was that China had 74 systems on the last TOP500 List and now it is clearly owns the

most number of installations, right after the U.S., as far as the TOP500 systems are concerned. And the whole world turned its attention to China at the end of last year, when it announced its brand new CPU, «ShenWei SW-3». The Sunway BlueLight MPP supercomputer, installed in Jinan, is composed of 8,700 ShenWei SW1600. The big shocker is that while the machine is nearly as fast as the Intel processors, it uses only one megawatt of energy, which could give the new machine and the Chinese a significant edge.

It was also quite amazing how Fujitsu came back into the game in a big way with K Computer at ISC'11. Some journalists came all the way from Japan to cover the TOP500 awarding ceremony.

From a program view point: Henry Markram's Conference Keynote on Simulating the Brain – The Next Decisive Years, Adisson Snell's Analyst Cross fire and the Panel on Energy Efficiency or Net Zero Carbon in 2020 were exciting happenings at ISC'11. We also held the HPC in Asia Workshop for the first time which attracted a large turnout among the Asian attendees.

In 2012, the SuperMUC installed at the Leibniz Research Center in Munich will be introduced, most likely as the new № 1 system in Europe.

I. L.: Has the governmental attitude to funding HPC and exascale projects changed recently?

H. M.: The best answer to this question would be to directly quote Ms. Neelie Kroes, the vice president of the European Commission. On February 15, she said the following in Brussels:

«High Performance Computing (HPC) is critical for industries that rely on precision and speed, such as automotive and aviation, and the health sector. Access to rapid simulations carried out by ever-improving super computers can be the difference between life and death; between new jobs and profits or bankruptcy».

«For these reasons, the Commission today sets out a plan for the EU to reverse its relative decline in HPC use and capabilities. Under this plan the EU will double its investment in HPC (from €630 million to €1.2 billion) and become home to computers that can perform Exaflops before 2020. Half of the investment would be for development and training and for new centres of excellence, creating thousands of jobs».

«Strengthening PRACE as the leading pan-European HPC e-infrastructure, pooling national and EU funds to service academic and industrial research, stimulating the market for HPC in Europe by supporting more acquisitions of HPC systems and services and faster uptake of HPC by industry and SMEs, establishing centres of excellence for software in scientific fields like energy, life-sciences and climate».

I. L.: Sometimes talks about the exascale era seem to be too far from real life. But maybe the opposite is true: is it too conservative to exclude Exascale from our horizon? Last month we were talking with Steve Scott about Zetascale. Are you going to touch the zeta theme?

H. M.: At ISC'12 we are going to address the importance of Exascale in different sessions, since our projections from the TOP500 lists demonstrate that we can expect the first Exaplops supercomputer in the year 2019. And seven years are not too far away. In the session «Heterogenous Architectures & Beyond» Exascale systems will be discussed since everybody is expecting that such systems will not be homogenous. We will have a highly interesting «Exascale chat» where Horst Simon from LBNL and University of Berkeley will chat with three different HPC authorities in the world about Exascale: Yutong Lu (National Defense University in China), Mark Seager from Intel, and Christian Bischof, Technical University Darmstadt. «Exascale Computing: Where are we?» is a session moderated

by Thomas Sterling from Indiana University. Speakers in this session will come from the U.S. Department of Energy, the industry and universities to discuss their view on Exascale.

At this year's conference we will not touch the topic Zetascale for the following reasons: If there are really 11 years between Exascale and Zetascale then we will not see the first Zetascale system before 2030. Discussions on HPC in 18 years from now are speculative and fictional from our point of view. While most of us rely on the fact that Exascale systems will be based on silicon technology, this will most likely not be true for Zetascale systems. Zetascale systems will most likely incorporate brand new technologies, e.g. quantum mechanics resp. other than that governed by Moore's Law.

In early March, it was reported that IBM researchers have taken a leap in computing by using quantum mechanics to harness the power of atoms and molecules, a move likely to lead to vast increases in speed and security of computers and other devices. So, let us wait and see. Answers were provided by: Wolfgang Gentzsch (General Chair ISC Cloud'12), Horst Gietl (Executive Consultant, ISC Events), Hans Meuer (General Chair ISC'12), Martin Meuer (Executive Director, ISC Events), Nages Sieslack (Public Relations Manager, ISC Events). ■■■

Special thanks to Wolfgang Gentzsch (General Chair ISC Cloud'12),

Horst Gietl (Executive Consultant, ISC Events),

Martin Meuer (Executive Director, ISC Events),

Nages Sieslack (Public Relations Manager, ISC Events).

Russia at ISC'2012



RSC Group

RSC Group (www.rscgroup.ru), leading Russian full cycle HPC solutions provider, developed RSC Tornado — cutting-edge energy-efficient scalable architecture with liquid cooling and industry-record PUE 1.06, offering solutions from mini Data Centers to large PFLOPS supercomputers with 60% power savings, outstanding 92% computing efficiency ratio by LINPACK in most custom flexible and green design, with own integrated software stack.

RSC Group will demonstrate at ISC'12 (booth #850) own HPC solutions on standard server boards (i.e. Intel) with liquid cooling, Intel® Xeon® E5-2600 processors, Intel® SSDs on board as well as deployed projects (the most energy-efficient in Russia by Green500), Intel® MIC support readiness with liquid cooling prototype.

Booth # 850

Moscow State University

Moscow State University (MSU) is the oldest and largest university in Russia. Founded in 1755, the University was named after its founder, Mikhail V. Lomonosov (1711–1765), a great Russian scientist. Its current rector is Academician Viktor A. Sadovnichy. The University has 40 faculties, 14 research institutes, 19 research centers. More than 40 000 students (graduate and postgraduate) and about 7000 undergraduates study at MSU, and over 5000 specialists do the refresher course here. More than 6000 professors and lecturers, and about 5 000 researchers work for the faculties and research institutes. Every year MSU enrolls about 4000 international students and postgraduates from all over the world. The flagship of MSU Supercomputing Center is 'Lomonosov' supercomputer with peak performance of 1.7 PFlops. 'Lomonosov' was created by 'T-Platforms' in 2009. After several upgrade stages it has now 5 104 x86-based compute nodes with Intel Xeon X5570/X5670 processors and 8 840 GPU compute nodes with NVIDIA X2070 cards. In all it has 52 168 CPU cores and 954 240 CUDA cores, with Linpack performance of 872.5 TFlops.

Computing part of "Lomonosov" occupies just 252 m², its power consumption is 2.6 MW. Other noticeable MSU supercomputers are SKIF MSU "Chebyshev" (2008, 1250 4-core Intel X5472 CPUs, 60 TFlops of peak performance, 47 TFlops of Linpack performance), HP "GraphIt!" (2010, 32 6-core Intel X5650 CPUs, 48 NVIDIA M2050 cards, 26.76 TFlops of peak performance, 11.98 TFlops of Linpack performance) and IBM Blue Gene/P (2008, 2048 4-core PowerPC CPUs, 27.9 TFlops of peak performance, 23.9 TFlops of Linpack performance).

Today MSU Supercomputing Center has over 550 users. Areas of users' research are hydro- and magnetohydrodynamics, quantum chemistry, seismic surveys, drug design, geology and materials science, nanotechnology, cryptography, ecology, astrophysics, engineering calculations, new materials design, and more.

In 2012 MSU take part in ISC exhibition for the 4rd time. MSU booth #125 at ISC'12 offers wide range of information about almost every MSU supercomputing activities: applications highlights, software development, education initiatives, and our supercomputing facilities.

Booth # 125

N. I. Lobachevsky State University of Nizhni Novgorod

N. I. Lobachevsky State University of Nizhni Novgorod (UNN) presented at ISC'12 is one of the leading Russian universities. UNN belongs to the set of Russian Research Universities that has been formed by the High-Priority National Competition Program. The University is renowned for innovation in research and higher education.

High performance computing (HPC) is among UNN's key activities. UNN Supercomputing Center features a computational hybrid cluster with peak performance of 117 TFlops (8th place in TOP50 list of leading Russian supercomputers). The cluster is running on Microsoft Windows operating system. The Center has computer hardware of many vendors, including Intel, AMD, IBM, Supermicro, NVIDIA, T-Platforms. At ISC'12 UNN is to present the best results of its HPC-based research.

Education. UNN has developed an advanced system of HPC education. At ISC'12 UNN is to present an overview of curriculum, courses and textbooks developed at the university. A brief presentation of the 'Internet University of Supercomputing Technologies' project (<http://www.hpcu.ru>) will be

given at the UNN booth. It is a joint effort together with the Moscow State University and the National Open University «Intuit».

More than a 150 HPC students were educated at UNN in 2011 with its student team winning the SC'2011 Student Cluster Competition ('Linpack Highest Performance'), in Seattle, USA.

Research. Over 10 research and applied projects are presented at the booth, including:

- *Management of high performance environments*
- *Decision support system for time-consuming optimization problems*
- *3D and stereo visualization of medical diagnostic data obtained by homographs of various type (CT, MRT, OCT, etc.)*
- *HPC in Medicine, Life sciences*
- *Scientific Visualization*
- *Machine Learning and Computer Vision, Robotics, etc.*

UNN's HPC research and educational activities are widely recognized. It also won the 'Informatics Europe Curriculum Best Practices Award' (2011) for Parallelism and Concurrency (jointly with MSU). UNN has also been awarded by Intel corporation with the Honorary Diploma for Outstanding Results in Training High-Qualified Experts in IT sphere.

Booth # 825

South Ural State University

South Ural State University (SUSU) is one of the largest universities in Russia more than 57 thousand students, 35 faculties, over 350 professors and 1500 PhDs. It has its own booth at ISC'2012.

SUSU has the status of National Research University. One of the key priorities of the SUSU is the development of supercomputer and grid technologies to address the problem of energy and resource savings. It is led by Leonid Sokolinsky, attending ISC'12. Leonidi is the Head of the world-class SUSU Supercomputer Simulation Laboratory running the 'SKIF-Aurora SUSU' supercomputer with the performance of 117 TFlops. The installation is among the top 4 Russian supercomputers (Russian TOP50), being the most energy-efficient HPC system in Russia according to Green500 thanks to the unique RSC Tornado liquid cooling technology.

Find out more about 'SKIF-Aurora SUSU' and see the 3D model of supercomputer at the center of SUSU exhibition booth. Powerful computing resources and high qualification enable SUSU staff to provide fundamental and applied research and innovation. At ISC'12 in Hamburg Supercomputer Simulation

Laboratory of SUSU will present more than 30 practical solutions, including:

- *Distributed Virtual Test Bed project. The aim is to develop a technology for automatic generation of the problem-oriented CAE/CAD Grid Services, enabling efficient usage of hardware and software resources. Proposed solution helps to generate a virtual test bed in the automatic mode and to carry out the virtual online experiments and multi-objective optimization in the Cloud using an Internet browser.*
- *Human Body project. The aim is to develop a technology of creation and usage of human body models for predictive modeling on supercomputer systems. This research includes the study of kinetics of deformation and destruction of human chest caused by a bullet impact on a bullet-proof vest and deformational changes at various types of vest fabric weave.*
- *Wide range of business and industry projects. The aim is to improve manufacturing process and develop new products for steel industry, machinery, defense, industrial construction, manufacturing, and thus enable close practical cooperation with the regional and federal enterprises.*

Booth # 112

An HPC workhorse

By Alexander Naumov

Any modern high-performance parallel cluster is a complex balance of many architecture decisions, taken long before the system goes online. While topology, interconnect, management, support infrastructure, and code optimization are paramount for successful cluster operation, the ubiquitous single x86 server continues to be the main HPC building block. The focus of this article is on a new volume HPC platform, recently developed by T-Platforms, an alternative to scale-out 'skinless' platforms, also known as 'twin' servers.

Published: #7 Autumn-2011

The T-Blade V-Class family is a modular x86 system, designed primarily for the HPC market. Following in the footsteps of the higher end TB2-XN and TB2-TL systems, T-Platforms developed a new V5000 enclosure, and two unique system boards to introduce 4 hot-plug compute modules: the V200S, V200F, V205S and the V205F. The V5000 supports up to ten standard 'S' or five double-width 'F' compute modules. The V200S is a high-end module with two Intel® Xeon® E5 2600 processors.

Its sibling, the V200F is a double-width, GPU-enabled module. On-die PCIe Gen. 3 controller provides ample bandwidth to accommodate two NVIDIA® Tesla™ M-class GPU accelerators (V200F-only) and optional on-board FDR InfiniBand/40 GbE VPI port. The system board also supports SAS 6Gb disk interface. The V205S is a value module with two AMD Opteron™ 6200 processors, providing a choice of eight 3.2GHz cores for frequency-dependent

environment or thirty-two 2.3GHz cores for multithreaded applications. To avoid PCIe Gen 2 bandwidth saturation, the V205F double-width module was designed to support a single NVIDIA® Tesla™ M-class GPU accelerator. Both modules come with SATA disks and optional on-board QDR InfiniBand/10GbE VPI® port. To increase system reliability, compute modules have no fans or cabled connections inside, except for the required GPU power cable. Even cold-

swap disk drives are connected directly to the motherboard using card-edge SATA or SAS connector, all of which makes V-Class modules stand closer to 'blade' architecture. Every compute module has 16 DIMM slots with two DIMMs available per channel and two integrated GbE ports. Customers can choose SAN/parallel NAS-based storage. They can also equip each compute module with up to two cold-swap 2.5" 1TB hard drives or SSDs to store temporary data or OS image. RAID levels 1/0 are supported. Compute modules with a convenient rear handle are installed at the back of V5000, with a latch mechanism clicking when the module is fully inserted. A lot of design effort went into the V5000 enclosure, which provides certain advantages over 'twin' servers. While the compute density equals most x86 'twin' servers, there is a component that sets the V-Class apart from other systems. V5000 features an integrated System's Management Controller (SMC): a 1U cold swap module with low-power ARM-based computer, 11-port Ethernet management switch with one front and one rear external GbE ports, serial port and an integrated KVM.

SMC provides centralized remote and local IPMI 2.0 monitoring and control of all compute module BMCs via an integrated Ethernet network. SMC also enables node/ OS-independent monitoring of the hardware sensors in the chassis. The IMU (Integrated Management Utility) is a web-based single system interface, which enables clients to assign static node BMC addresses, observe the system's health status, set alarm thresholds, update or roll back firmware, and troubleshoot nodes with no reliance on command-line interface, or on HPC or enterprise management software. While each compute module has power, reset and unit ID controls at the back, there are also extended status LEDs and controls located on the system front for both the compute modules and the chassis. The enclosure is installed in industry-standard 19-inch rack cabinets. Special

attention should be paid to instance where PDUs on each side of the rack cabinet can potentially block the extraction of compute modules in slots #1 and #10. When dozens or even hundreds of V-Class systems are deployed, users can dramatically reduce their Ethernet cable clutter by utilizing SMCs external GbE management port instead of dedicated node connections to consolidate platform and cluster management.

V5000 is an air-cooled enclosure with three in-line cooling modules (6 fans), installed in the front section of the chassis, known as the «cold zone». As compute modules have no fans inside, architects designed a fully passive low-profile midplane to ensure sufficient, direct front-to-back air stream. Reliability is reinforced with female-type connectors on the midplane to avoid any contact damage in the chassis on module insertion. Customers can easily replace simple card-edge connectors in compute modules in the field. While the system has still to pass homologation for the Northern American market, it already supports both high-line and low-line power input. There are four highly efficient '80Plus Platinum' 1600W power supplies at the back of V5000. They rely on 200-240VAC /50 Hz, 1-phase input and support 110VAC for most hardware configurations. Both cooling modules and power supplies are hot swappable and provide N+1 redundancy.

The T-Blade V-Class system sports the density of 2 servers per 1U of rack space, yet compared to the 'twin' design, there is a higher power supply and cooling fan consolidation rate to make the system even more power efficient. Several low power configurations with low voltage components are available with three redundant power supplies instead of four.

A straightforward V5000 midplane, connecting all system modules together, routes just power, ground, control and monitoring signals. As there are no integrated data or

compute network switches, every compute module has external ports to connect to Ethernet and to InfiniBand interconnect switches. In most small- and medium-sized scenarios, where the number of ports is less or equal to 684, clusters often use one central InfiniBand switch; therefore, any built-in switches or pass-thru modules can make the network infrastructure more expensive, complicated or oversubscribed. Thus, V-Class does not limit the customer's choice of the preferred network equipment vendor, supporting virtually any topology just like most scale out systems in HPC and cloud computing today.

As a result, a highly reliable, medium density V-Class, based on industry standard components, is compatible with many HVAC or power delivery and backup subsystems in place today, and can be quickly deployed in existing HPC and cloud computing centers. To serve customers better, T-Platforms plans to introduce several customer-ready V-Class solutions, based on rack cabinets with an optional cold door. A 42U cold door solution is expected to support up to five V5000 enclosures with 50 nodes, head node, switches, and support infrastructure with modular managed PDUs within a 22 kW range.

V-Class is to support the new compute module types, future processors and accelerators, including upcoming AMD 'Piledriver' and Intel® 'Ivy Bridge' processors, and NVIDIA® 'Kepler' accelerators. With updates to the SMC to improve power efficiency and control granularity and a planned lifecycle through 2016, the T-Blade V-Class is a worthwhile investment for both HPC and higher end 'cloud' environments.

The new HPC 'workhorse' is a reasonably priced energy-efficient platform, supporting 'mix and match' of various Intel- and AMD-based nodes and GPU enabled configurations. Perfectly positioned in the market between 'twin' servers and blade systems, it covers a wide spectrum of HPC and cloud applications today.

Supercomputer for Superjet

By A.A. Ryabov, LLC «Sarovsky Engineering Center», M.S. Litvinov, V.P. Popov, CJSC «Sukhoi Civil Aircraft»

Illustrated by Vladimir Kamayev, Alexander Zhelonkin

Published: #6 Summer-2011



Modern passenger aircrafts must adhere to the highest national and international safety requirements. To meet these levels regarding various possible emergency situations, the developer of a new aviation technology shall analyze all the design solutions that affect safety at the design stage and confirm compliance of such design with items of the certification basis through the aircraft certification procedure.

The conventional technology of proving the reliability and safety of an aircraft and its parts requires mock-up modeling and the conducting of full-scale experiments. However, with the ever-increasing safety requirements and expanding possible accident scenarios, the scope of required mock-up models and tests increases considerably, which inevitably leads to increased costs and elongates time for new product development. In these circumstances, there is a very promising approach to address such safety concerns, which is based on the optimal combination of computer technology and computer modeling experiments. Being skillfully applied, this approach has undeniable advantages, allowing to receive a large amount of information needed for comprehensive analysis of engineering solutions, confirming the reliability of products, with significantly reducing development time and cost. Safety concerns of Superjet-100 CJSC «Sukhoi Civil Aircraft» (Moscow) has developed a new medium-haul airliner Superjet-100. To certify the aircraft, it is necessary to prove its reliability and safety in possible emergency situations, such as landing

gear tread collapse during takeoff and landing, as well as landing with the landing gear being partially extended. In case of the tread collapse, a jet of compressed gas and fluttering reinforced rubber fragments which have been throwout of can cause significant deformations of some structural elements, or even destroy them. Therefore it is necessary to study the possible consequences of such accident to prove the safety of the aircraft. Since the destruction of the tread can happen in any place and impact on different elements of the aircraft, located in the possible impact area, it is necessary to analyze dozens of scenarios of such accident. In case of an emergency landing with the partially extended landing gear, it is necessary to prove to the certification agency that the deformation and possible partial destruction of some elements of the aircraft will not lead to catastrophic consequences. It is obvious that in both cases the reliability checks by means of full-scale dynamic testing involves significant financial costs. So, contemporary technology of computer simulation, becomes a very promising alternative to solving

complicated safety problems of aircraft engineering.

Computer technology

The computer technology is based on numerical methods for solving problems in mathematical physics and continuum mechanics. At present the finite-volume methods are widely applied in fluid and gas mechanics, the finite element methods - in the mechanics of deformable solids. These methods are based on different sampling schemes by space and time of non-linear integral and differential equations describing physical processes. Therefore, to solve practical problems, which are usually characterized by complicated shape of studied elements, discrete (mesh based) models are developed, describing all the special geometric features of products. The accuracy of the solution depends on the quality of the sampling, as well as the adequacy of the physical models used to describe the physical processes. Detailed discrete models, including all the structural elements that influence the behavior of the system, with space resolution from millimeter fractions to

several millimeters are required for the numerical solution of safety concerns of the Superjet-100 with high accuracy. A comparison of the required time and space grid with the size of the actual design shows that the discrete models shall have millions of cells, finite elements, resulting in tens of millions of equations. The solution of such problems within a reasonable time with the required accuracy can be only accomplished by supercomputing technologies based on the use of clusters with hundreds or more processors. To analyze the consequences of the tread collapse, one needs conjugate solutions of gas dynamics problems relating to the gas jets outflow and impact on the various structural elements, along with the solution of the problems of dynamic deformation of these elements from the gas jet pulse pressure, as well as from the impact of the tread fragment. Research of emergency landing comes down to solving the problems of dynamic deformation and possible destruction of aircraft elements when landing under specified conditions. In order to solve the complex problems of unsteady gas dynamics and dynamic deformation of aircraft structures, LLC «Sarovsky Engineering Center» applies well-known in the world licensed software - STAR-CCM + and LS-DYNA. Generation of discrete models of large dimension is based on special tools that are built into STAR-CCM +, which enables clients to work with the original CAD-information of the developer.

Reliability of numerical modeling

The most important concern regarding the quality of the computer technology-based modeling is to confirm the validity and accuracy of numerical results. The adequacy of computer modeling is proved by comparing the numerical solutions with the results of a series of model experiments, which are similar to the physical processes occurring in

the studied design of the aircraft. In this case, the model experiments are represented by a series of tests, where the gas jet and tread fragment impact on the flat plate, simulating the landing gear compartment wall, as well as on fragments of the actual aircraft pipeline and cabling, provided by the developer. Model experiments correspond to emergency conditions in aircraft by specific size of structural elements, mechanical properties of materials and thermo physical properties of the gas, as well as by the levels of the initial pressure and velocity. The experiments results are empirical data of the gas jet pulse pressure jet and the deformation of model structural elements. This empirical data forms the basis for verification of the developed computer models, the adequacy of which is confirmed by the proximity of the numerical results and empirical data. After the reliability of used computer technology is confirmed, numerical studies of Superjet-100 safety in these emergency situations are carried out.

Computer-aided analysis of accidents

For the numerical modeling of emergency tread collapse by CAD-models, the CJSC «Sukhoi Civil Aircraft» developed a series of highly accurate digital models of the landing gear compartments and deformation models of the given structural elements. In total, more than 30 possible scenarios for this emergency situation were considered. Specific models of internal volume filled with gas and deformable elements being developed for each scenario. The main problem that is solved by a specialist in computer modeling is generating high-quality digital models, carefully describing geometry of real structural elements, without any «simplifications» of geometry and design features, while proving the adequacy. Usually «simplification» often involves no less effort than solving the very problem.

For the development mesh models of a complex tread gas volumes, LLC «Sarovsky Engineering Center» uses a very efficient software tools of STAR-CCM + - Surface Wrapper and Surface Remesher, allowing to detect all errors of the basis geometric model, quickly repairing them in automatic mode and creating an adequate discrete model in a short time. Polyhedral mesh consists of 2.5 to 7.2 million cells with a detailed resolution of the prismatic boundary layers have been developed for solving problems. To generate finite element deformation models of separate aircraft units in the landing gear compartment, with up to 1.2 million hexagonal finite elements, as well as load-bearing elements such as the main landing gear, with 2.4 million finite elements – a pro-STAR software was used. Workstations with RAM not less than 16 Gb are used for the development of such discrete models. Translation pressure of the gas jet to the surface of deformable elements is performed on a specially designed converter that allows to automatically transfer the load from STAR-CCM + to LS-DYNA without mesh modification. STAR-CCM + software is used to simulate gas dynamics. It is based on an implicit scheme in conjunction with a combination of methods for solving linear algebraic equation: multigrid method (AMG) and the preconditioned conjugate gradients method. Dynamic deformation is studied by an explicit time integration scheme, implemented in LS-DYNA. To solve the problems, it used up to 96 cores of 128-core computing HP cluster with processors Intel Xeon CPU 5160 3Ghz and each node RAM being 32 Gb. Numerical modeling shows that both STAR-CCM+ and LS-Dyna have a high scalability of 75-85%. Conjugated numerical calculations of gas outflow from the tread gap, dynamic deformation modeling of the elements under the impact of the jet and tread fragments allowed with high accuracy to estimate the strength of all given assemblies, enhancement some structural elements, confirm the safety

of the aircraft in case of such accident. To study the emergency landing, the developer provided a «static» finite element model of the aircraft, detailed solid models of units and assemblies, mechanical properties of materials and conditions for landing. Based on these data, a detailed «dynamic» computational model of the aircraft was developed. This model includes the built-in primary elements and tie-down fitting, units and assemblies of the design, affecting the safety of the aircraft, takes into account the contact kinematic relations and elastoplastic nature of the deformation of materials (which is necessary for high-precision computational analysis and prediction of the effects of dynamic deformation in the emergency landing conditions). Established models allowed to simulate the four given scenarios of landing with not extended and partially extended landing gear. As a result of computer modeling, the accuracy and reliability of the technical solutions that ensure the safety of the aircraft Superjet-100 in an emergency landing, was proved.

Research results

Using the state-of-the-art computer technologies based on cluster computing resources, LLC «Sarovsky Engineering Center» for the short time received a large amount of information about the behavior of critical assemblies and elements of the aircraft in the given emergency conditions in case of the tread collapse and landing with partially extended landing gear. Basing on the results of calculations, the CJSC «Sukhoi Civil Aircraft» finalized the design of separate elements of the aircraft, improving its safety. The executed works helped to saved time and money while proving the reliability of Superjet-100 according to the important items of the certification basis, which contributed to the successful certification of the new passenger aircraft confirming its compliance with the contemporary safety requirements. ■■■



Predicting the future: TOP500 2018, what it will be like?

Editorial interview

Published: #1 Spring-2010

The editor of our magazine Igor Levshin talked to Jack Dongarra (University of Tennessee), guru of the HPC industry, author of the famous Linpack test and co-author of the TOP500 list and one of the initiators of the draft of International Exascale Software Project (IESP) – an international project to develop software for the coming Era of Exa Computing.

Igor Levshin: When will future of the exa computers come?

Jack Dongarra: There are U.S. supercomputing industry development projects. It is estimated that exaflops performance will be achieved by 2018 (plus or minus 3 years). At that time the supercomputer will be able to perform one billion threads at the same time. It will have 10-100 million cores.

I. L.: Will this scale require the revision of Linpack tests?

J. D.: A review will be needed, but primarily due to changes in the architecture of machines. Exa computers will have hybrid architecture.

I. L.: Do you think that the hybrid systems will be the mainstream of the high-performance computing by this time?

J. D.: The next generation of systems will be equipped with multi-core

processors working together with different kinds of accelerators, also multi-core. In the future GPU cores will be integrated with traditional CPU on one chip, so the processor will have a high floating point performance.

I. L.: What are the prospects for the use of FPGA in the supercomputing industry?

J. D.: They are good for certain tasks, but for general purpose calculations they are often not suitable. FPGA must be programmed, but it is much harder to create a regular program. Upgrading to the latest version, using a different more modern algorithm – all this is more suited to traditional programming. FPGAs are good for applications that are «eternal». Then it is efficient and cost effective. FPGA will ideally be used for a long time when it comes to scientific computing, and specific tasks.

I. L.: «Mass» processors from Intel and AMD are widely used in the supercomputers. In your opinion, do «elite» processors such as POWER7 or processors of the SPARC architecture which Fujitsu supplies for RIKEN have a future?

J. D.: At the level of national laboratories, not all is determined by the laws of mass-market processors. In any case, the division of consumer processors and processors for high performance computing does not reflect the effect of trends in this market as a whole. Cray now often uses the consumer processors. But at the same time Cray has its own technology of very fast interconnect, so these systems can not be attributed to the consumer ones, and Cray is definitely among the industry leaders. As POWER7 is a very powerful processor, IBM will do everything to ensure that POWER7 is in demand

not only in the scientific computing market, but also in other markets. Of course, much is now determined by the consumer computers market, but the midrange systems will be powerful computers, and it will have an impact on the processor market as a whole.

I. L.: Will there be the large international projects, analogs of CERN physicists in the era of exa computing?

J. D.: Yes, this is a great opportunity for the development of the industry. We are currently working on International Exascale Software Project (IESP). There are many members of the «Big Eight»: US, UK, Japan, Russia, Germany, France and Canada. The goal is the development of software for scientific computing. This program was launched two months ago and the developers have to give preliminary proposals in order to have a chance of getting subsidies until May.

I. L.: The topic of power of supercomputers is widely discussed nowadays. Cooling is also one of the hottest topics. Does it sound reasonable to build future data centers somewhere in the north?

J. D.: Exaflops systems which is discussed in the US have to consume not more than 20 MW. Otherwise, an entire power plant will be needed. And the project will not be able to meet the limit of \$200 million. It is also important to remember about the network bandwidth. In addition there are also security issues too.

This question becomes very relevant since there is data center equipment that costs many millions of dollars. There are many problems not related to the site location however. The memory is a very critical and expensive component of a supercomputer. The memory access bandwidth is a problem that must be solved in the nearest future.

I. L.: What will improve the reliability of a supercomputer?

J. D.: This is a very important question, because it will be necessary to monitor the failure of hardware, software bugs and the temperature.

The main thing is to learn how to predict where failure may occur and not simply react to the crash occurred. But it is necessary to be able to solve the problems on several levels: to predict, to be able to react to the failure and to be able to correct the consequences of a failure.

I. L.: Is revolution in hardware or software of supercomputers possible in the future?

J. D.: I think that all of the major changes will occur in evolutionary rather than revolutionary ways.

I. L.: How will the landscape of supercomputers market change in the future?

J. D.: Look at the TOP500 list. You will see that 60% of supercomputers are used for commercial purposes. The biggest machines are still focused on the research goals, but most of them are used to enhance the competitiveness of commercial companies: telecommunications financial, engineering companies, etc.

I. L.: By the way, is it reasonable to refer to Google's giant clusters as specific kinds of supercomputers?

J. D.: They have really powerful centers. If they would run Linpack tests, we could judge about the productivity of their clusters. But the performance tests require hours of work rather than minutes or seconds. Not all companies can afford it. In addition, for companies such as Google, their own IT capabilities are an essential tool of competition, so they do not disclose this information.

I. L.: What do you think about the future of quantum, optical, and other calculations? How far away is this?

J. D.: All this is guesswork. It is not easy to predict the development of the industry, even at 10 years, and we are talking about decades here. All of this will develop: optical calculations (I'm talking about optical logical elements, not interconnects) and bio computing, but it will not be a real basis of calculations soon. I think that there will not be any radical changes in hardware components in the next 30 years.



Jack Dongarra – one of the initiators of the draft of International Exascale Software Project (IESP) – an international project to develop software for the coming era of Exa Computing (the other co-founder of IESP – Pete Beckman from the Argonne National Laboratory).



I. L.: What could Russia's contribution to the advancement of the global computer industry be? Maybe this will be HPC-software?

J. D.: International cooperation is just such a path. That's how you can turn the Russian experts in the creation of advanced technology in our industry. Many ideas and mathematical approaches have come and are coming from Russia. For example linear algebra that is one of the most important thing in supercomputing. A lot of important ideas have Russian origin.

I. L.: Is experience of the U.S. in this area applicable for Russia?

J. D.: The applications areas like energy, environment, digital video and animation, aerospace and biotechnology are not very different in various countries. In the US, Russia and other countries supercomputers are used around the same goals: to research, educational purposes and for business. ■■■



Smart chips do more with less.

V-Class V200S and V200F compute nodes

- A variety of Intel® Xeon® E5 2600 processor-based configurations
- Up to 256GB of node memory and up to 2TB of local storage
- Ten V200S nodes with 20 CPU's per chassis
- Five V200F nodes with 10 CPUs and 10 GPUs per chassis
- Energy-efficient design
- System management controller for consolidated node management



**Powerful.
Intelligent.**

T-PLATFORMS



- Custom HPC hardware development
- HPC system software development
- Storage expertise
- 'Turnkey' HPC system deployment
- Modelling, simulations and resource-intensive computing on demand

T-Platforms is a global supercomputer developer and a supplier of the full range of solutions and services for high performance computing.

Modern software and hardware solutions for aircraft engine applications

By: R.K. Gazizov, A.V. Yuldashev, A.M. Yamileva,
Ufa State Aviation Technical University (USATU)

Published: #7 Autumn-2011

It seems that the world of parallel computing will never stop to change. Over the past decade, HPC clusters filled world supercomputing ratings and pushed aside all alternative architectures. Parallelism has gone beyond the confines of the supercomputer world and has been explicitly appeared in the personal computers. The increasing number of cores in general-purpose central processors (CPUs) is no longer surprising. Recent years were marked by the new trend. Graphics processors (GPUs) began to be used for acceleration of general-purpose computing.

So how the latest architectural innovations affect the solving speed of real-world problems? This is what we tried to find out by comparing efficiency of simulating the process of linear friction welding with MPI version of ANSYS Mechanical package

using two computing systems of different generations. This testing was carried out at USATU HPC cluster based on Intel Xeon Clovertown processors and hybrid computing system based on more recent CPUs (Intel Xeon Nehalem) and NVIDIA

Tesla GPUs. Applied software vendors are doing their best via adaption of software packages to the new computer architectures. Thus ANSYS, the well known finite element analysis package, began to support simulation

acceleration on NVIDIA GPUs in 2010. Starting from version 13.0, this feature appeared in shared memory solvers of ANSYS Mechanical module for thermal and structural analysis. The release of version 14.0 allowed once to use GPU acceleration of ANSYS Mechanical simulations on distributed memory systems as well. However, currently only one GPU on compute node can be used. During our research we compared both versions of ANSYS and two different hardware platforms mentioned before. Linear friction welding (LFW) becomes a crucial technology for coupling difficult-to-weld materials. This technology has been successfully implemented in the manufacture of aircraft gas turbine engines blisks based on the "buildup" process, as an alternative to milling parts from solid blanks. LFW technology allows to repair damaged blades instead of whole blisk replacement.

In Russia, the LFW technology for developing the aircraft engines of the next generation is being implemented at JSC "The Ufa Engine Industrial Association" (UMPO). Finding out the best welding modes depending on the geometry of welded surfaces and used materials requires prior mathematical and computer simulation. Such research is conducted by employees of UMPO and USATU within the framework of the project "Development of technologies and industrial production of assemblies and blades of gas turbine engines with lightweight high-strength constructions for the aircraft engines of new generation"¹ using high-performance computing resources of the university.

Friction welding is a kind of pressure welding, in which heating is generated due to the friction caused by the movement of welded parts relative to each other. Since the heat produced

by friction is localized within thin surface layers of welded metal parts, i.e. exactly where it is needed, such welding process has a number of important advantages.

- High performance – the complete welding cycle lasts about several seconds.
- Low energy consumption - in the process of welding the material does not reach the melting point, being heated only within a small area near the contact surface of the welded parts.
- High quality of weld and the stability of welded joints quality.
- Ability to weld metals and alloys in various combinations, and less strict requirements for the preparation of pre-welded surfaces.
- Hygiene of the process (no UV radiation, harmful gas emissions, hot metal splashes, etc.).

In the linear friction welding, heat is generated during the reciprocating motion of welded parts with a frequency of about 60 Hz and an amplitude of to 3 mm being compressed in order to form a tight

contact. There are four stages in the process of linear friction welding. At the initial stage, welded products are brought into contact under pressure and their relative motion begins which is accompanied by a deterioration of roughness. At the transition stage, heating and elastic deformation of work pieces take place. Having reached the yield point, the equilibrium phase comes in which is characterized by axial shortening. Plastic material is displaced from the contact area (Fig. 2). At the final stage, the mechanical motion is finished and additional forging pressure is applied to the samples in order to form a welded joint.

It is important to note that only the computer simulation of the elastic stage of linear friction welding (which actually takes about 0.2 seconds) is calculated for about a week if performed in a sequential mode. Therefore, complete process simulation without usage of parallel computing is out of question. The high complexity of the simulation is concerned particular with following.



Fig. 1. A non-detachable connection of turbine buckets with turbine disc as an example of an application of linear friction welding process

¹ This project has been implemented within the framework of Regulation No. 218 of the Russian government of 9 April 2010 "On measures of state support for the development of cooperation of Russian higher education institutions and organizations implementing complex projects for high-tech production".



Fig. 2. A sample of a linear friction welding joint

Large stresses in the contact area lead to intensive heat generation and high temperature gradients requiring fine-grained mesh for simulation. A coupled (structural and thermal) problem is solved. The transience of simulated process necessitates the choice of a small (10-4 ... 10-5 s) time step for the convergence of computational methods. Also the dependence of mechanical and thermo-physical material properties from the temperature (physical nonlinearity) and the change of loads and boundary conditions in time (structural nonlinearity) are taken into account. Thus, during our comparative tests, we have investigated the simulation time of LFW elastic stage for steel samples on various computing platforms (Table 1) performed with MPI version of ANSYS Mechanical software

package. The geometric model consists of two rectangular samples placed one on another (Fig. 3). The model contains 18,432 elements with the average size of 0.55 mm. At each time step a system of about 300,000 equations is solved by using the SPARSE method. Initially let us consider scaling of parallel simulation with MPI version of ANSYS v13 and v14 on IBM HS21 compute nodes with x86 processors (Fig. 4). Speedup is given relatively to the time of sequential computation on one processor core in ANSYS v13. The maximum 25x speedup was achieved on 64 cores and no further core count increase was possible due to the limited number of licenses. One can see that with a fixed number of parallel processes and a fixed number of processor cores respectively computation goes faster when using

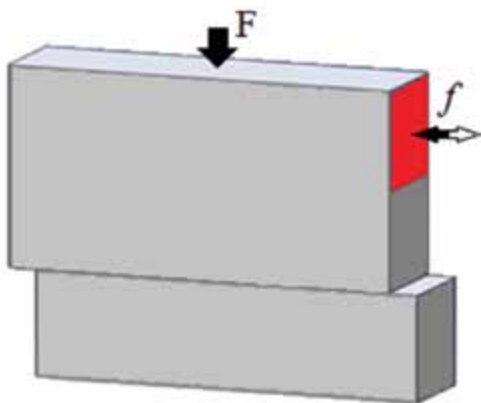


Fig. 3. Geometric layout

as few cores at the node as possible due to lack of processes contention for shared resources. In addition performance can be improved by 5% when running simulation in ANSYS v14 comparing with v13. Now let us consider the results of testing on a hybrid node with graphics processors. Speedup is measured relative to the same time as earlier. More recent Intel processors show visible performance improvement. Performance of computation on 8 cores within one node increased by more than 2.5 times relative to the previous results. The maximum speedup obtained on CPUs was achieved on 12 cores when running new ANSYS version (v14, CPU, see Fig. 5). Here usage of updated ANSYS v14 provides a significant (up to 1.5 times) performance improvement compared to the

Table 1. Specifications of computational nodes applied for testing

	Computational nodes, installation year	Peak node performance	Number of nodes used	Processors	Accelerators	Random Access Memory (RAM)
1	IBM HS21 comm. environment InfiniBand 10 Gbps, 2007	74.56 GFlops	up to 64	2x Intel Xeon 5345 2.33GHz	none	8-64 GB, DDRII 667 MHz, ECC, FB-DIMM
2	IBM iDataPlex dx360 M3, 2011 r.	140.64 GFlops (2xCPU) + 1 030 GFlops (2xGPU) = 1 170,64 GFlops	1	2x Intel Xeon 5670 2.93 GHz	GPU NVIDIA Tesla M2050	48 GB, DDR3 1 333 MHz, ECC

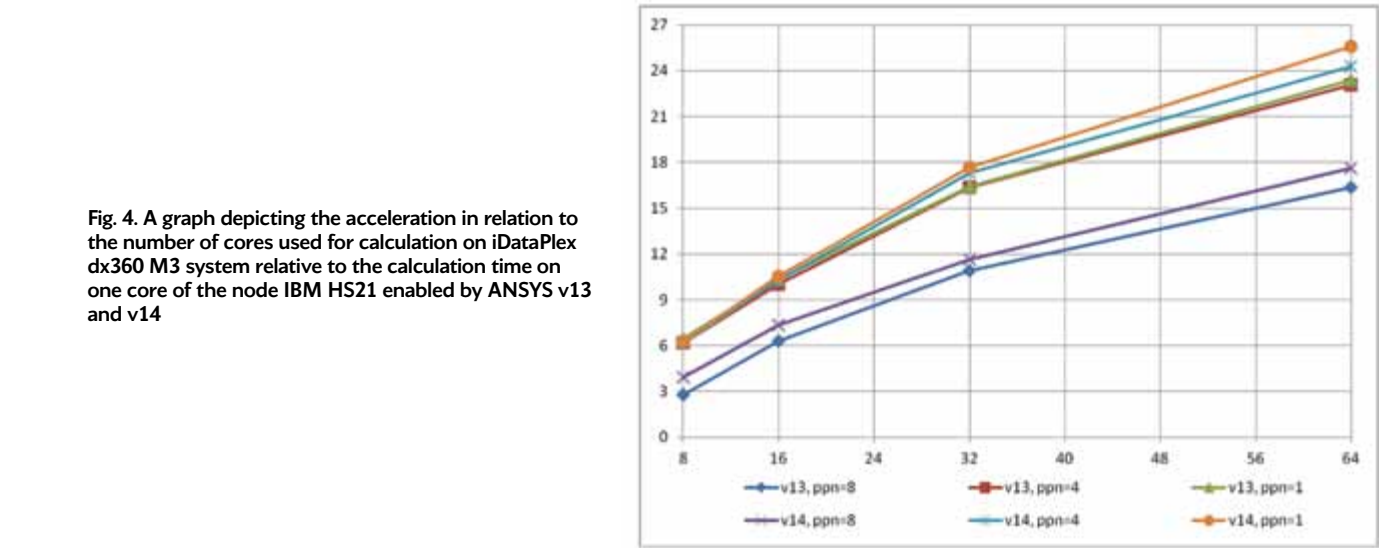


Fig. 4. A graph depicting the acceleration in relation to the number of cores used for calculation on iDataPlex dx360 M3 system relative to the calculation time on one core of the node IBM HS21 enabled by ANSYS v13 and v14

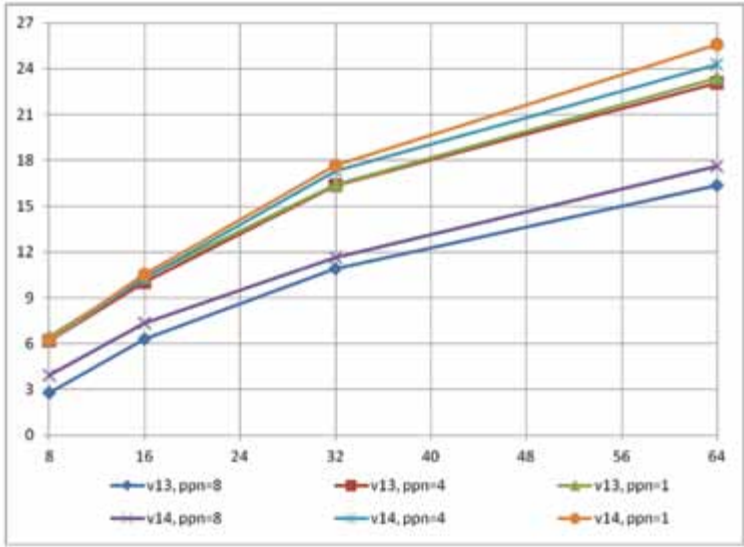


Fig. 5. A graph depicting dependence of simulation speedup on iDataPlex dx360 M3 from number of cores relative to the simulation time using ANSYS v13 on one core of IBM HS21 node

previous version. In addition to higher computing performance on CPUs, the MPI version of the ANSYS Mechanical v14 software package allows to use one graphics processor per node in order to accelerate computations. It allowed to achieve a record-breaking performance of hybrid compute node (v14, CPU + GPU on Fig. 5). Simulation time on 8 CPU cores and one GPU is close to that on 8 HS21 nodes of HPC cluster. Moreover, we should note that in the last case the second graphics

processor of the hybrid node remains idle. Therefore, when massive computations are needed it seems reasonable to run two jobs on a node, say, each on four CPU cores and one GPU. Related experiments (v14, CPU + GPU (2x) on Fig. 5) have shown that such workload insignificantly decrease performance of each simulation allowing to achieve almost double performance in massive computations. Consequently new hybrid system may replace around 12 rather than 8 HS21

nodes of previous generation for examined application! Generally speaking, in our case the minimum time of thermo-structural simulation concerned with linear friction welding process has been achieved on multiple HS21 cluster nodes based on the Intel Xeon Clovertown processors. However, the performance of a hybrid system was found impressive. This means that modern hardware and software solutions are capable of conducting simulation processes on a brand new level. ■■■

Swallow in the clouds

Editorial interview

(Tsubame in Japanese means «swallow»)

Published: #2 Summer-2010

Igor Levshin interviews Satoshi Matsuoka, the ideologist of the Tsubame project, professor of Tokyo University of Technology.

Igor Levshin: Matsuoka-san, let's start with Japan. You have glorious supercomputer traditions...

Satoshi Matsuoka: Yes, Japan has always been strong in the computer business. NEC, Hitachi and Fujitsu had competed internationally with companies such as giants as IBM or DEC. But there were also interesting companies known predominantly for consumer electronics that had made efforts to be visible in the higher-end computer world. Sony, for example, once made workstations with RISC-processors. With regard to these three giants, NEC, Hitachi and Fujitsu, now they play a major role in the global market, not only in Japan. NEC has good positions in Europe; Fujitsu owns the computer business Siemens and Hitachi is among the world's leading storage systems. If you only talk about supercomputers, then again, Japan has a long history. Particularly the Japanese had been strong in supercomputers based on mainframe-like technologies, and thereby always competing with IBM. In general, only

Japanese supercomputer companies were able to compete with Cray in the design of vector supercomputers. Then there was famed vector Earth Simulator. Now the 10 Petaflops K Computer of the next-generation will be created. This is the priority national project. But lately, as you probably noticed, there has been a slowdown in growth and significant drop in the overall market shares.

I. L.: Do you mean the Global crisis?

S. M.: It's not just that. I think that the assimilation of new technologies is slow because of company-internal competition, the complex relationships within the company. Leading Japanese companies have their own design of high-performance processors with a complex architecture. Presently, the rapid growth is usually associated with many-core CPUs – Intel, AMD, and now with NVIDIA GPUs. To abandon their projects in favour of these that are have leverage in the mass market is not easy. The glorious traditions and their own processors are not always



conducive to progress.

I. L.: Tell us about your brainchild Tsubame 2.0.

S. M.: There are now only three months until the start of the production. There are more than 4 thousand NVIDIA Tesla GPU's in a Tsubame 2.0. and the

peak performance will reach approximately 2.4 petaflops. We have been conducting research on GPUs for supercomputing for a long time. Currently our Tsubame1 supercomputer has been retrofitted with 600 NVIDIA GPUs in addition to the existing helping 10,000 CPU cores.. We use it for production-level experiments; we are working out the technologies to ready them for production. We are engaged in research on supercomputers with the GPU with NVIDIA and Microsoft. In general, we have a lot invested in GPU as a promising architecture. It's not easy but it is and interesting, because we have several thousand users, and they have to explain what the GPU is and why you need it, why they should think about this and about what next generation of supercomputers will be equipped with GPUs. So, it makes sense to start it right now. So we teach users how to migrate to new architectures. Our Tsubame 2.0 will be small in size and highly efficient machine. NEC will be the main integrator. All compute nodes are produced by HP (we have been working with HP for more than a year) and some service nodes are also made by HP. There will be the new HP motherboards, a new HP chassis, designed specifically for HPC; they are not like normal servers, in Tsubame 2.0 almost everything is new. It is very energy efficient; it takes up very little space and is very comfortable to control it. It will have a very high bandwidth of «processor-memory» path, of «input-output» and of the network.

I. L.: What parameters are most important for Tsubame 2?

S. M.: We'll have high performance regarding the LINPACK benchmark, but it is not our main goal. The system, according to our calculations, will work effectively with a variety of different applications: with hydrodynamic applications such as packages of weather forecasting, which require large bandwidth. In addition, these applications will be

considered very effective, including in terms of energy consumption and the ratio of «performance to power consumption» will be very high. I even think that we could become number one in the Green500.

I. L.: Will Tsubame 2.0 suite cloud computing? What do you think about the prospects of HPC clouds?

S. M.: In the future clouds will add HPC-resources. Another thing is that many now believe the concept might not be as good as some thought. Concerning the software, hardware and the cost of computing – all of this so far had not met expectations. And there are no cloud data centers in TOP500 yet. To date, supercomputers are too expensive for these applications. Now the cloud centers mainly work with very little «heavy» large-scale parallel applications. But I think that by its very nature the demands on supercomputing are not so different from those on the servers in the cloud data center. So the economy of large-scale projects will win and that the convergence is inevitable. There are dozens of supercomputer centers in Japan. But the problem is that existing centers already have a limit of development: they exhaust the possibilities of power supply to them; their sizes are limited by the DC floorspace. It hinders the growth. A need for supercomputing in Japan is far ahead of the pace of construction of such centers. Existing supercomputer centers will not solve this problem, and universities and other large consumers of HPC will seek services elsewhere. A new data center for cloud computing could be the answer. But we need modern, efficient supercomputers. Now there's a discussion of a joint project with Microsoft. We need virtualization – which goes hand in hand with cloud. This is a great project that will involve hundreds of developers. We will study how to balance the load within the network and how to work effectively with MS Azure. It is important for the company. From the other side we have the opportunity to work with

Microsoft software for HPC. Thus we divide the costs of the project. We also will have Linux and we will communicate with the creators of other cloud platforms.

I. L.: What is the situation with HPC education in Japan?

S. M.: We don't just have a simple data center; we also have a HPC research lab at our institute which is high end. In general, there is a good tradition of HPC culture and training in our university. It's in many ways similar to MIT. We help students with contacts and provide easy access to supercomputing resources so the students are getting used to the HPC from the beginning. There are also good HPC departments in other Japanese universities. They will use our supercomputers too.

I. L.: When did you start your career in computing? What experiences are the most interesting?

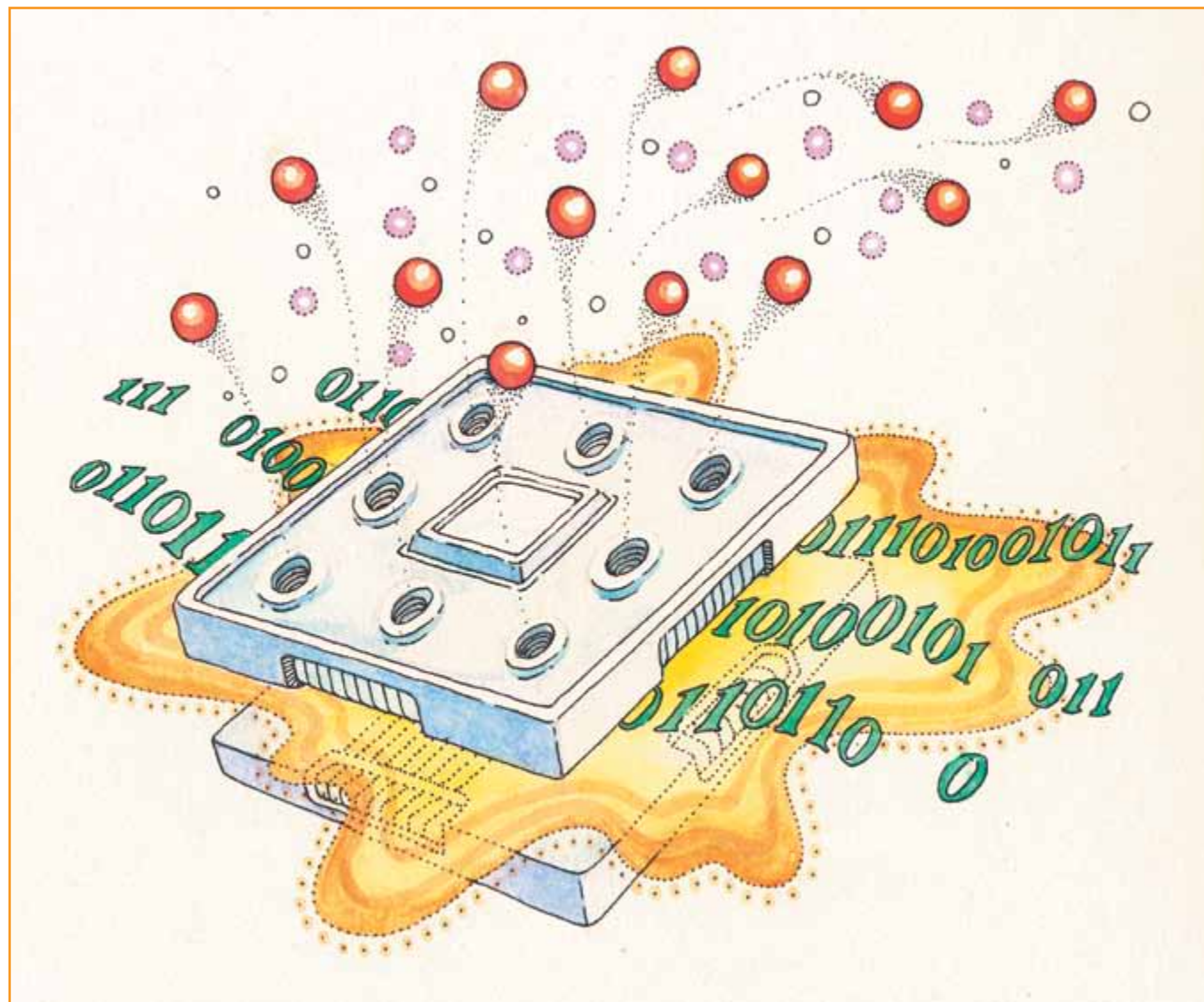
S. M.: I first sat down at computer when I was 14 years old. It was in the 1970's when the first mass microprocessors appeared. I became a programmer, I programmed Nintendo and I earned a part-time living. I met with HPC in university and at the time we used parallel machines. Later on, in the early 90's we mostly used Intel and Fujitsu models. They were bulky machines then – with more than a hundred processors. So my experience with parallel computing I suppose, is over 20 years. But my most valuable experience came when we began to make clusters at Tokyo Institute of Technology in my laboratory. We started to develop supercomputers ourselves. So it was a real school! The processors with multiple cores appeared, so we had to constantly think about the systems architecture and how to analyze them. By the way, I always thought that the software is a very important part of supercomputer technology. We started to build clusters – a process which we are still doing today. In total we've now been working with them here for more than 13 years. ■■■

Cooling systems of the future

By V.L. Kovalev, A.N. Yakunchikov
Illustrated by Alexander Zhelonkin

Microelectronic components keep on to be reduced in size, and the amount of energy dissipated by their cooling systems, has been steadily increasing. In this regard, the issue of electronic components cooling is quite acute.

Published: #7 Autumn-2011



It is assumed that the future cooling system will be a system of micro- and nanochannels, penetrating the electronic component. Such channels will give way to the circulation of cooling liquid or gas. Prototypes of such cooling systems have already appeared in the research institutes in the United States. The cooling device consists of two basic elements: an emitter and a collector. The emitter is a system of charged needles with spikes of small diameter, creating charged air particles. The collector creates a chip-cooling flow at the expense of the electric field impact on the ionized medium. The flow can be also created by a «micro-pump» – oscillating wall of the channel. In most cases the flow of gases and liquids is studied on the basis of the macroscopic approach, when the medium is considered to be a continuous one [2, 3], thus not taking into account its discrete (molecular) structure. Such description is valid in cases where the studied volume contains a sufficiently large number of molecules, so the medium can be considered continuous. However, when studying the flow in micro- and nanochannels, modeling physical and chemical processes in gases and on surfaces, in some cases it is reasonable to use a microscopic approach

based on molecular dynamics and direct computational modeling [4-8]. With this approach the corpuscular structure of the gas is taken into account, the positions and velocity of molecules are determined at any specific time, and the macroscopic values are identified with the mean of the corresponding molecular values. Intuitively, everyone understands that the smaller the diameter of the cooling system channel, the greater the density of their allocation in cooled volume and, consequently, the greater is the ratio of the cooling system channel wall square to the cooled volume. Therefore, the reducing of the cooling system channels size results in heat-transfer enhancement. But the heat transfer flow is proportional not only to the wall surface square, but also to the heat-transfer coefficient, which depends on the characteristics of gas-dynamic flow. And since, as mentioned above, the equations of gas dynamics are not suitable for describing the flow and heat transfer in microchannel, then the heat transfer coefficient should be determined in a different way. If the heat transfer coefficient decreases sharply when using micro- and nanochannels, then it makes no sense to produce such a demanding cooling

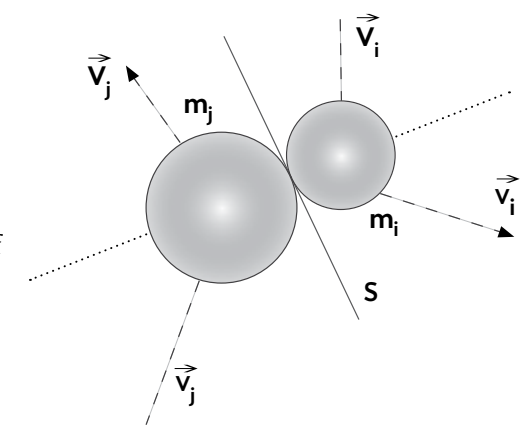


Fig. 2. Model of particles interaction

system. Taking to account such reasons, scientific minds of our time, specializing in molecular dynamics, decided to determine the behavior of the heat transfer coefficient value using numerical methods with the help of the supercomputer power. How did they do it (and still do, as there is a great variety of gases and liquids)? This will be discussed further.

Mathematical modeling of flow in micro- and nanochannels

How are the flows in micro- and nanochannels studied? What models and techniques are used? At the first glance it's all pretty simple and does not extend beyond the school curriculum in physics. The flow of heat-conducting perfect gas between two plates, located at distance L_y (Fig. 1), is studied. Usually, to reduce the dimensionality of the task, the flow is considered to be two-dimensional, and the flow range symmetric to a plane equidistant from both walls of the channel. The task is solved by means of direct computational modeling. This means that the gas is not seen as a continuum, but rather as a collection of moving molecules, with the true gas flow being described by a huge number of particles, the change of

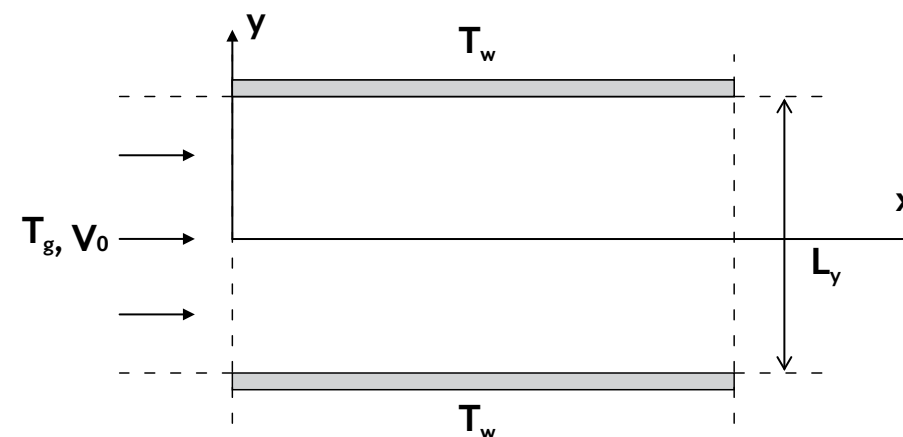


Fig. 1. Study of the flow of heat-conducting perfect gas

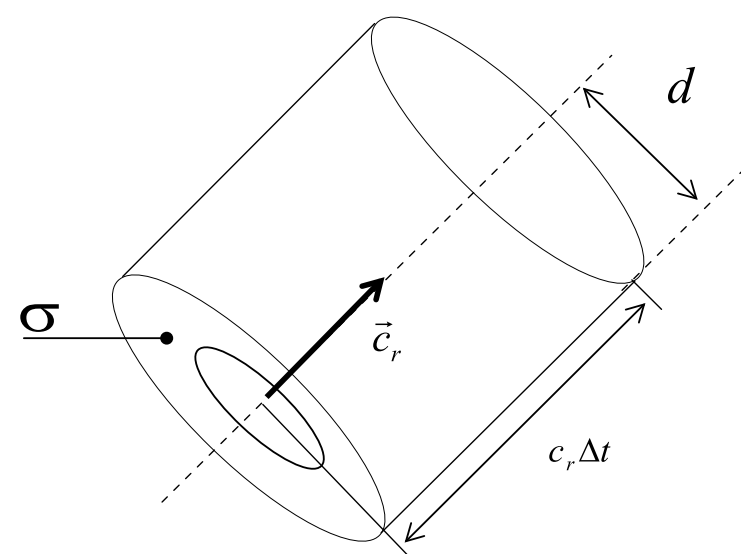


Fig. 3

coordinates, velocities, and properties of which are determined over time by intermolecular interaction and interaction with the channel borders. It is the number of particles that determines the need to employ supercomputers for the direct computational modeling.

$$\frac{d\vec{v}_i}{dt} = F_i(\vec{x}_i, \vec{v}_i, t),$$

where \vec{x}_i and $\vec{v}_i = \frac{d\vec{x}_i}{dt}$

The equation of the i -th particle motion from Newton's second law can be written as follows coordinates and velocity of the i -th particle, and the right parts of these equations F_i are given in accordance with the chosen model of particle interaction. As to our task, typically one uses one of the most common patterns of interaction – the molecules are represented as impenetrable spheres with diameter d , and it is believed that the interaction takes place only due to collisions (hard-sphere model). The molecular velocities after their collision were calculated by the

$$m_i \vec{v}_{\perp i} + m_j \vec{v}_{\perp j} = m_i \vec{v}'_{\perp i} + m_j \vec{v}'_{\perp j}$$

$$m_i \vec{v}_{\perp i}^2 + m_j \vec{v}_{\perp j}^2 = m_i \vec{v}'_{\perp i}^2 + m_j \vec{v}'_{\perp j}^2$$

where m_i and $\vec{v}_{\perp i}$

velocities before their collision using the law of conservation of momentum and energy conservation law: mass of the i -th particle and its velocity, perpendicular to the plane S (Fig. 2).

To probabilistic approach is often used to reduce the computational efforts in the description of the collision [5].

The more complex diffusion model is used to describe the interaction of gas with the surface. And in this case it is assumed that the velocities of each molecule after reflection do not depend on their individual falling rates and are distributed according to the balanced Maxwellian distribution function in the half-space of velocities, where the molecular velocity vector is directed from the surface. The distribution corresponds to the temperature of the surface. Scheme of the modeled domain is shown in Fig. 4. The flow is considered symmetrical, so the modeling is carried out on one side of the symmetry plane. To do this, a condition of particle plane reflection is laid.

In domain A the following flow is organized: gas is cooled to temperature $T_g = 0,9T_0$ and reaches the mean velocity v_0 . The plate in this domain has temperature $T_w = T_g = 0,9T_0$. Domain B is the studied one, where the plate has temperature $T_w = T_0$.

Results

Fig. 5 shows the distribution of the dimensionless velocity V/V_{av} in the channel cross section (V_{av} – the mean velocity over the cross section) depending on Knudsen number $K_n = \lambda / L_y$, which determines the validity of the assumption of the medium continuity. With small values of K_n , when the description of the flow in the framework

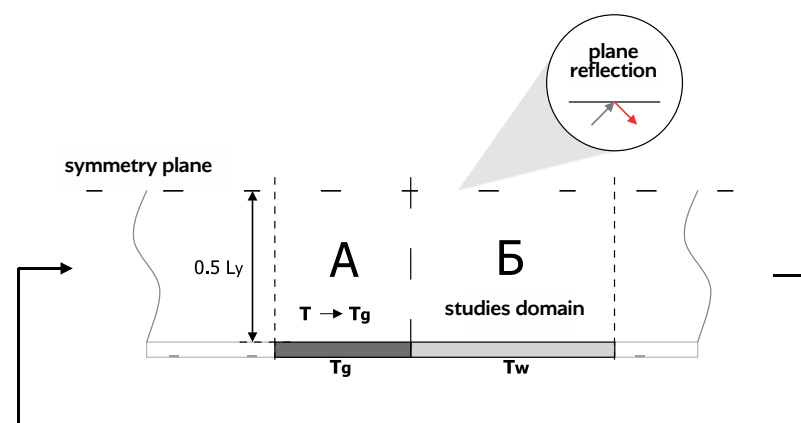


Fig. 4

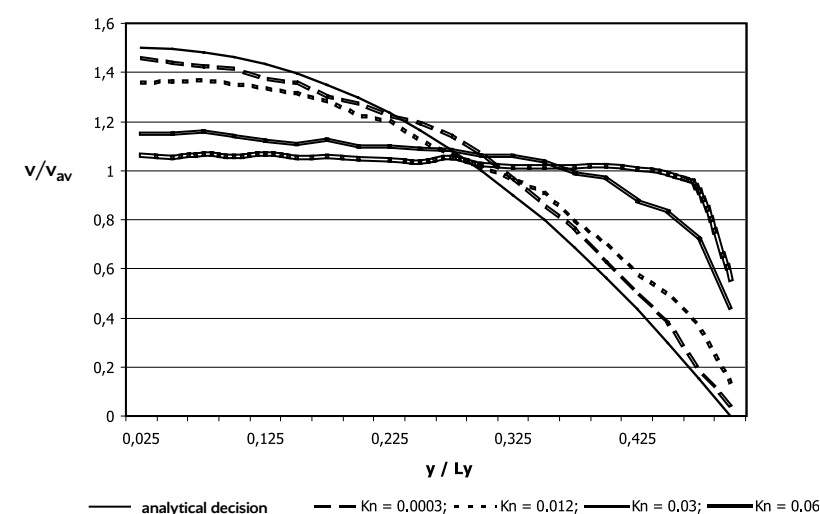


Fig. 5. Velocity profiles depending on different Knudsen numbers

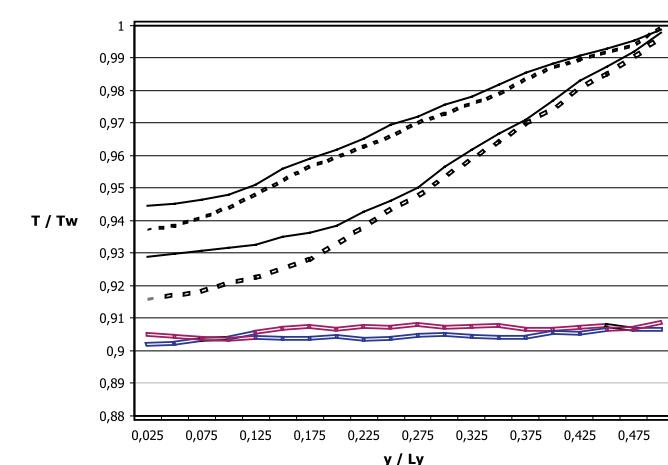


Fig. 6. Profile of temperature and some channel cross-sections

of continuous medium is valid, a parabolic velocity profile is obtained and the calculations are the same as the solution of Navier-Stokes equations for selected flow conditions. With sufficiently large Knudsen values K_n , where the mean free paths are comparable with the linear dimensions of the task, the velocity profile is almost straightened, and the wall impact is implied only in the immediate vicinity.

Fig. 6 shows the distribution of the dimensionless temperature T/T_w in several channel cross sections with various Knudsen values. The calculations were carried out with $T_w = T_0$, $T_g = 0.9T_0$, $K_n = 0.006$, and $K_n = 0.03$. As one can see, gas is heated faster in the channel of smaller width. Near the surface at the initial heating area the dimensionless temperature gradient with $K_n = 0.006$, and $K_n = 0.03$ is equal to 0.222 and 0.133, respectively.

But in dimensional coordinates, the temperature gradient (and hence the heat transfer flow from the surface) with $K_n = 0.03$ is almost three times more than with $K_n = 0.006$.

Fig. 7, with $K_n = 0.006$, shows the typical distribution of the density,

velocity and temperature in the studied domain. The numerical values of the parameters should be estimated by the color saturation and a scale of values appropriate for such domain. As follows from these results, our concerns about the sharp decline in the heat transfer coefficient in case of

increasing of Knudsen value (reducing the size of the channel) have not been justified. Moreover, in case of reduced channels the heat transfer coefficient increases, which gives reason to call micro- and nanochannels systems the cooling systems of the future.

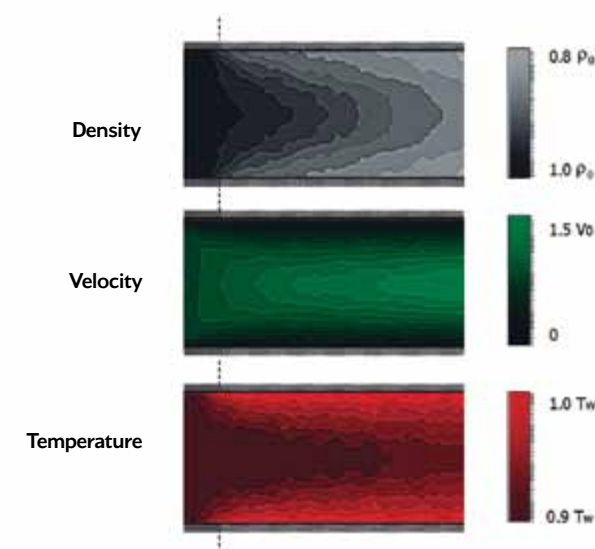
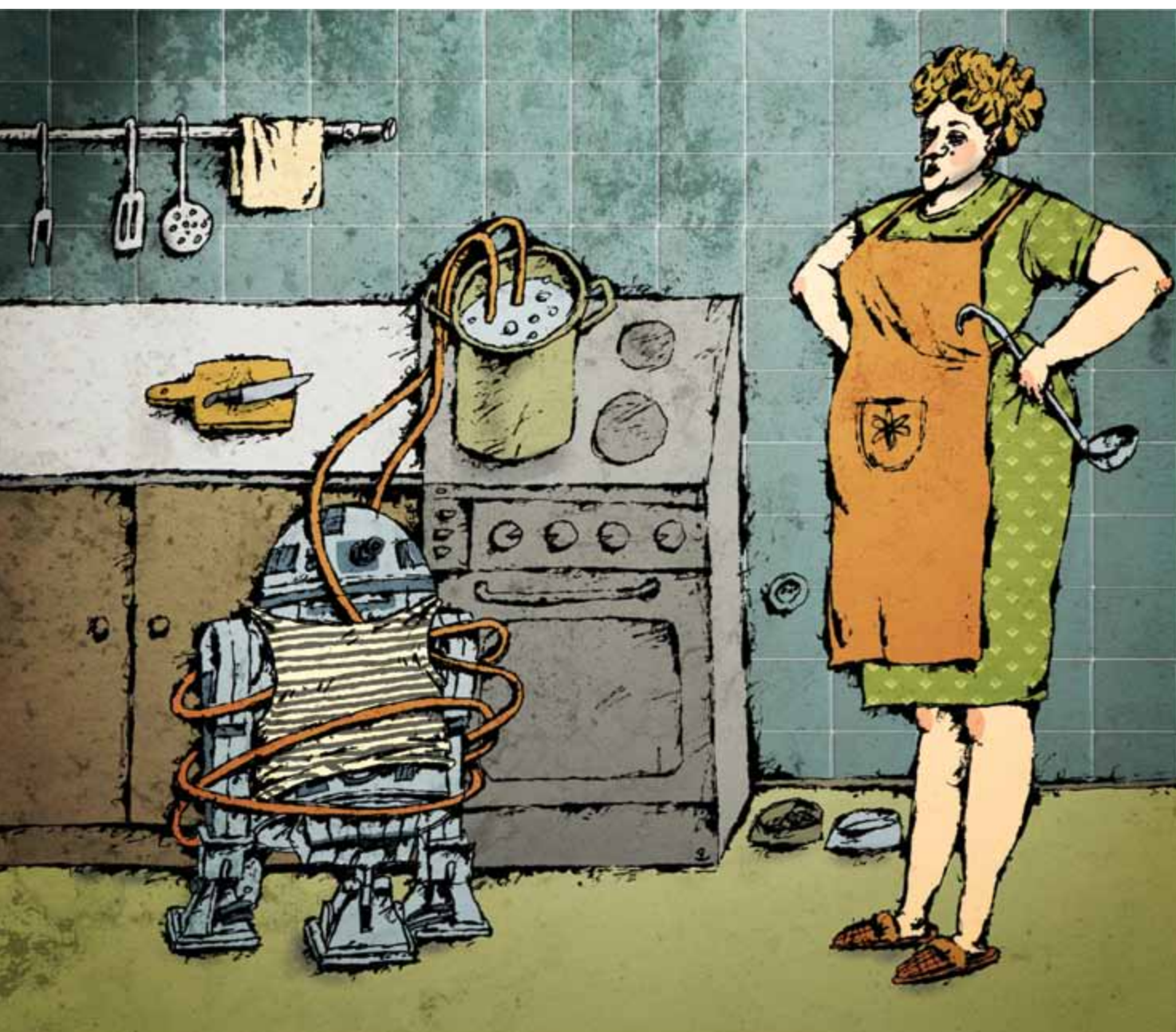


Fig. 7 Distribution of density, velocity and temperature

Water-cooled machines: another false start?

By Igor Obukhov
Illustrated by Vladimir Kamayev

Published: #9 Spring-2012



Everyone knows that computers can be efficiently and economically cooled by water. But everyone knows, that when someone says «everyone knows» it usually means something controversial and/or untested.

Thus, what are the advantages of water cooling?

1. More efficient cooling of processors.
2. Compact size compared to air cooling systems.
3. Possibility of cooling by "hot water", significantly increasing PUE (Power Usage Effectiveness).
4. Possibility of utilization of heat generated by computers for heating or cooling (using adsorption chillers).

However, water cooling has disadvantages, which should also be kept in mind:

1. A more complex service.
2. Water in the immediate proximity to power components.
3. Some of the system components could not be cooled by water.

Of course, these minuses are not exactly minuses, they depend more on implementation details. For example, the current rating leader of TOP500 supercomputers, the K Computer, built in Japan, has only water-cooled processors, usually representing about half of total energy consumed by computing node. Turing back to merits, it seems that these advantages are obvious. But let's look at them more closely.

1. A more efficient cooling of processors.

It is difficult to argue, but only because all mass-produced processors are perfectly cooled using air cooling. CPU temperature can be maintained at the desired level. What is to argue? Yes, there are experiments on the

direct water cooling of silicone. There are attempts to develop "multi layer" chips where the coolant is fed into the «sandwich». But, apparently the result of these experiments will be closed systems, which will dissipate the heat on the lid of the processor. Otherwise, there will be a need to build a large circuit with very clear water – where in addition to special filtering systems and the need to apply special pipes, fittings and valves as any dust, mechanical and chemical contamination will cause problems. Given that with the supply of water directly into the silicone pressure drop will be substantially higher than in ordinary water systems, it will require the use of special high-pressure pumps, which further complicates the systems and increases power consumption. In general, the greater the efficiency seems to be there. But what is the criterion? If we assess the ability of the cooling system to ensure the CPU temperature within acceptable limits, water and air cooling... has the same efficiency.

2. Compact size compared to air-cooled systems.

In a nutshell – it's not. In fact, the problem to fit a water cooling system, even in the usual 1-unit dual CPU computing node case – is not trivial. In addition, there are quite thick pipes with a thick layer of insulation that should be connected directly to each of the compute nodes. And

in fact regarding power density air cooling systems are still in the lead. And yes, of course, if it is all developed from the ground up: rack, chassis, motherboard, power supply system of all compute nodes and the cooling system of computational nodes – the water cooled system can be more compact than an equivalent performance production air-cooled blade system's. But on the other hand, if the same approach to the issue is applied to increase the power density of air cooled blade system from scratch, most likely it will be possible to achieve even higher power density.

3. The possibility of cooling by «hot water», thus significantly increasing PUE (energy efficiency).

This is perhaps the largest of the possible advantages of water-cooled systems. One problem – most of the currently existing systems, water-cooled computers still use fairly low water temperatures, and installations without the chiller in the cooling system has so far been rare. Causing increased PUE. On the other hand, developers of conventional systems designed air conditioners that promise cooling system PUE below 1.1 for many regions. More about it will follow.

4. The possibility of using heat generated by computers, for heating or cooling (using adsorption chillers).

At present it seems to be the greatest advantage of water cooling systems. It will be really cool to turn 2, 3 or even 10 MW of heat exhausted by a supercomputer for something useful: i.e. heating greenhouses or to support a nearby residential area.

The problem is that supercomputers can not be a guaranteed source of heat. There may be some malfunctions, scheduled maintenance or software upgrade – if this is the case then the computer heat production reduces. After 3-5 years of operation, supercomputers can become out of date and most likely means it will be shut down for about 3-6 months for an upgrade. In other words – not reliable.

Consequently, there must be a certain back-up heating system, operating in case of problems with computer heat production. And here's the «crux of the matter». First, the cost of conventional heating systems with «traditional» energy sources is added to the cost of heating systems using supercomputer heat. Second, the «normal» heating system, if not used, requires additional maintenance. Third, its startup process can take

up to several hours. With outdoor temperature –20°C, a few hours without heating will greatly cool down any home.

Well, perhaps the most important thing: (yes, of course, it depends on energy supplier) very often energy suppliers impose very severe penalties on consumers not spending a certain quota agreed at heat source connection. Very often these fines are much higher than the cost of energy spent for heating! In addition, there are a number of purely technical nuances. For example,



Fig. 2. K-Computer rack. Computing nodes are mainly cooled by water, but memory, system board, power supply units and other components still need air cooling.

many heating systems can't use water with temperatures below 55°C at room radiators; otherwise there is a risk of bacteria growth (including the famous legionella). Given that the supercomputer's cooling circuit cannot be directly connected to residential heating circuit, it means use of the heat exchanger. This, in turn, means that the temperature in the cooling circuit of supercomputer should be at least 5°C higher. So the output of supercomputer heat sink, considering losses in the pipes, should have temperature of more than 65°C, meaning it can't be used for some CPUs.

There is an alternative: one can utilize energy of water with lower temperature, using heat pumps. But in this case, still there are additional costs for the construction of the system, maintenance and energy consumed for heat pump operation. Another interesting thing is that the manufacturers of air-conditioners are not idle either. Over the past few years, a number of systems using adiabatic or indirect evaporative cooling appeared.

What principle do they follow? Typically, it is an air to air heat exchanger, sprayed with water when the ambient temperature does not allow to cool air in the internal circuit down to 25°C required in the computer room. Or by simply using «outdoor» air feeding through filters and «evaporation chamber», where water is sprayed, thus cooling and humidifying the air. As an additional



Fig.1. System IBM Aquasar with water cooling. As it requires heat exchange modules, pumps, and space for water conduits, the power density is lower than for the air cooled IBM blade system.

Over the past few years, a number of systems using adiabatic or indirect evaporative cooling appeared. What principle do they follow? Typically, it is an air-air heat exchanger, sprayed with water when the ambient temperature does not allow to cool air in the internal circuit down to 25°C required in the engine room. Or simply using «outdoor» air feeding through filters and «evaporation chamber», where water is sprayed, thus cooling and humidifying air

«source of cold» in especially hot days, a compressor-condenser (DX) unit is usually used, similar to an ordinary air conditioner.

This block cools air if adiabatic cooling is not sufficient, and also could be used to reduce humidity.

This approach allows us to obtain an average PUE below 1.1 for the large part of Europe.

Piping for water cooling system with «warm water» is simple and effective: the «inner» loop filled with water, the heat exchanger; the «external» loop filled with a glycol solution – an external heat exchanger (dry cooler). But redundant pumps should be used for inner and outer circuits.

You could make a system with one circuit, but then cooling system should be filled with a glycol solution, which would require a higher flow of coolant, because the heat capacity of glycol is much lower than that of water - and in combination with the higher viscosity of the solution of ethylene glycol, higher pressure is needed in the system.

As a result, the cooling of one megawatt of compute nodes by «hot water» require up to 100 kW of power for circulation pumps and fans of external heat exchangers. Plus the megawatt of heat cooled by water usually means a 100 to 200 kW of heat dissipated through the walls of cabinets, plus heat of storage systems and communication equipment that require the use of air conditioners. Thus, it appears that the actual PUE already shown on a real working air-cooled systems is very close to PUE, which could be theoretically reached by the water cooling system.

At the same time, the maintenance of air cooling system is usually easier and less expensive, cooling system is less complicated, and the use of hot/cold aisle containment allow us to reach very high power densities.

As a result, the water cooling systems that started to appear in the wake of the struggle for power efficiency of supercomputers, it seems, once again lost the battle before it started.



More work must be done in less time

Editorial interview

Published: #3 Autumn-2010

A conversation with inventor of Beowulf clustering technology, Thomas Sterling, a Professor at the University of Louisiana. Interviewed by Vladimir Voevodin and Igor Levshin.

Igor Levshin: Do you remember the time when Beowulf took shape? Looking back, what seems naive, and what was a foresight?

Thomas Sterling: The original Beowulf project, which was launched in 1993, was purely experimental. We tested the hypothesis that real-life scientific problems can be solved by using a system composed of components typical for consumer market, such as single-processor nodes, local network and software models plus message transmission tools, which came up by the time. Our project kicked off under the influence of another, earlier work. There were farms of workstations, which enabled the experiments that studied the performance of clusters of specific applications. The work was sponsored by NASA HPC Program and its goal was to find ways how to achieve cheaper gigaflops performance while solving real problems, at the time a system of such capacity would cost about \$1million. We did not see our ideas as naive, because we were professionals who empirically tested the hypothesis. We did not regard ourselves as prophets either. At this time, our Beowulf was criticized from all sides due to

various, sometimes opposite reasons. Supercomputer community, which has traditionally relied on huge machines (such as the Big Iron) just hated us.

Well, if we were prophets, our prophecy would have boiled down to development of the industry of large numbers, the mass market and consumer devices and software, rather than relying on more expensive special architectures. If you look at the current list of TOP500, you will see that we have guessed the formula of success: Beowulf is a combination of today's most common network, that is Ethernet, a cluster of consumer devices, which is today's most common architecture, the most usable Linux software, the most popular programming model built on transmission of communications and ubiquitous processor architecture x86. Not a bad guess and note we got it right from the first try!

I. L.: Do you think exa architectures will develop by way of evolution or revolution?

T. S.: This is a key question and very hard to answer. Both sides have enjoyed the support of the industry's top professionals. My opinion is that for many purposes it is enough to use



clouds with exa resources that could be built on the evolved traditional petaflops methods and structures. But if you take specific exa tasks, scientific and research applications that must operate efficiently, be fairly scalable, be convenient and reliable, then it would take a revolution of the systems' architecture. Exa systems of the future will be very different from traditional architectures with distributed memory and message transmission. Because they will maintain Global Address Space (GAS) as opposed to currently used

distributed memory; putting to work multithreaded systems, rather than multiprocessor architecture, enabled by light synchronization mechanisms for rapid transfer of control, rather than on the principle of global barriers; they will have dynamic resource management systems, rather than rely only on static allocation performed during compilation; they will be more likely to rely on a powerful run-time system than the slow OS services, process data as soon as they are available by moving the code to the data, which is quite the opposite to what is typical now in a message passing model: the movement of data to the place where they will be processed sometime in the future, rely on micro-architecture built on the ideas of data-flow, rather than on processor cores, which intensively use speculative performance of instructions – this is the way to saving energy. Also they will more likely use Embedded Memory Processors – (EMP) in order to reduce delays and increase the sampling rate of data-intensive memory processes, even in case of bad temporal locality, rather than rely on slow and inefficient – in terms of power consumption – migration of data through cache hierarchy; use light micro-control points directly within the memory, rather than harsh restart from hard drives for reliable recovery in case of failures, and use common methods for programming various types of cores within heterogeneous structures.

For these and other reasons I am sure that we are now approaching the phase transition to the HPC and that the differences of exa era system architectures from the current ones will be revolutionary.

Vladimir Voevodin: Will these revolutionary concepts require new institutions? What will it take to put these ideas into practice?

T. S.: Again, it all depends on whether the changes will be revolutionary. I think yes, but many clever people disagree and their opinion also should

be taken into account.

I think that we have already entered the 6th phase of the HPC evolution, which means that you need a leap to a brand new model of computing. I prefer the model which I call ParalleX. It was designed in order to handle the most critical problems such as hunger (when the system is short of work which could be effectively distributed), overheads (which are determined by effectively treated minimum granularity of parallelism), latency (the number of cycles required for remote access or addressing a query) and conflicts (e.g. delays that are also measured in cycles that are spent in waiting for access to shared logical and physical resources).

In ParalleX and similar models, these problems are solved with a set of means, which mainly have been used before, but now they have to be put together for the first time ever – both PGAS and multithreading with hardware context switching, and the of model calculations on data availability, plus lightweight synchronization mechanisms. All these jointly are able to deal with the degradation of performance. For tasks associated with processing graphs, the model of calculations, based on «data availability», allows to build the calculation process so as to discover and use the parallelism inherent to metadata.

Given such mechanism, management acquires dynamic rather than static nature; run-time information will help distribute the jobs in the queue, which can't be predicted during compilation time.

But there's a trap: it is not so easy to increase efficiency, because more work (in the run-time) must be done in less time.

Many institutions aim to expand the existing practical capabilities as far as possible. And it makes sense. Their depths are unlikely to bring to life any new technologies based on new models of performance, at least not until the end of this decade, in which developers of supporting

system software, programming environments and architectures have to fit, in order to meet the needs of the future exa scope of computing. This task should be performed by other institutions, or even by new ones that have been, especially set up for this purpose and they must be maintained in order to be able to perform such work. But this is risky: nobody today can guarantee that the chosen way is correct, until we have approached the goal.

V. V.: Last year, at the Supercomputing Conference in Hamburg, you talked about «Sterling Point», which is an insurmountable barrier for supercomputing performance. Do you still assess the capacity of this point at 64 Exaflops?


What is the outlook for technology, which could help surpass this barrier?

T. S.: «Sterling Point» (thank you for reminding it) is just the limit of performance (of course, approximate), which by my estimates, will never be surpassed in principle by any general-purpose system built on binary logic Boolean gates and discrete floating point calculations.

Simply saying, we will never reach any zetaflops. I estimate this limit to stand at 64 Exaflops, but it will be more accurate to say that this limit is found somewhere in the range of 32 to 128 Exaflops – this depends on what kind of price we are willing to pay in terms of size, power consumption, cost and reliability.

This limit has fundamental grounds such as the speed of light, Boltzmann's constant and the structure of atomic nucleus. There are other boundary conditions that result from the theories of information, noise and some others.

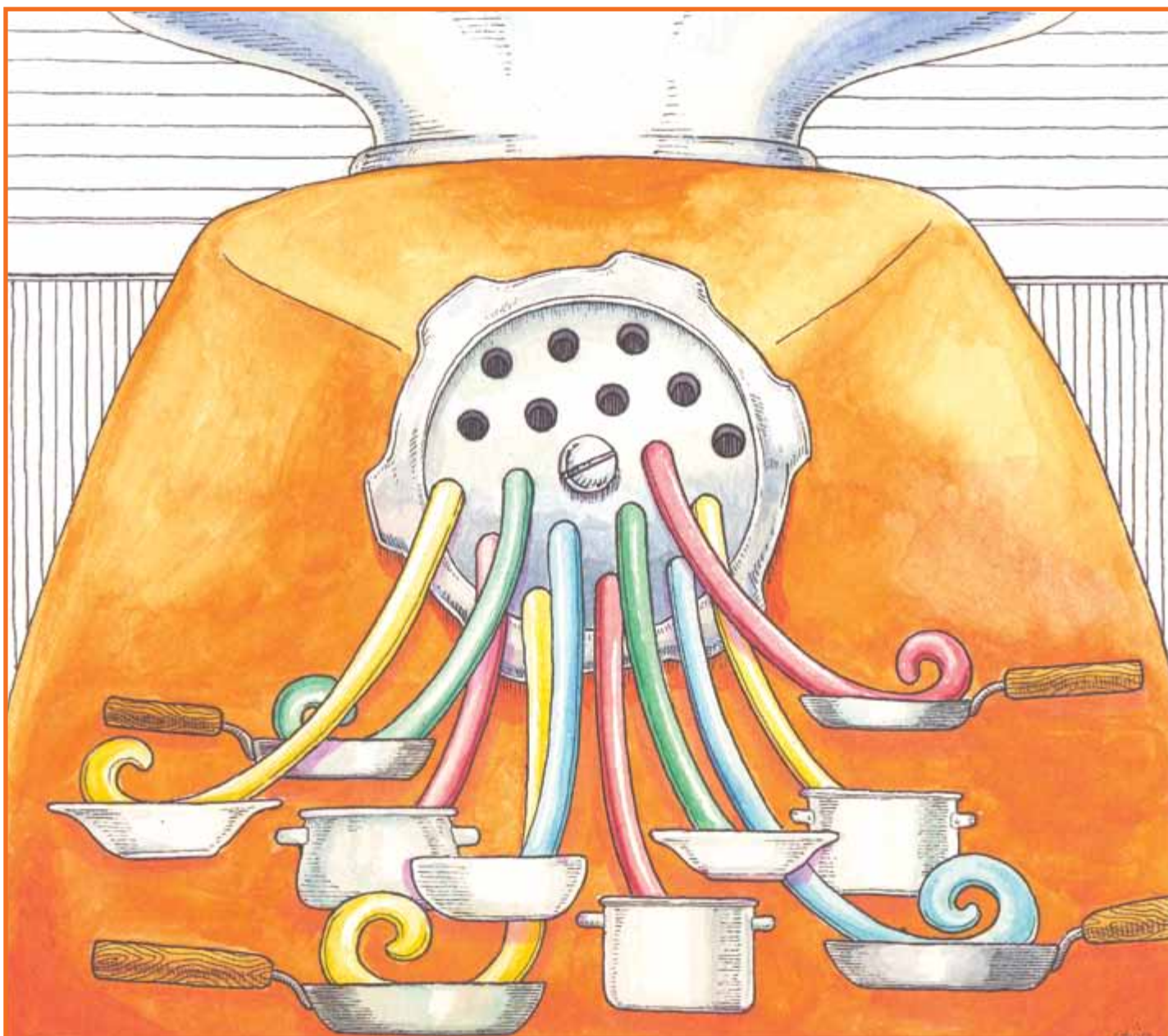
But I am very apprehensive because there are other paradigms of calculations that have been already investigated or will be discovered and they will not be limited by the «Sterling Point».

Perhaps it will be a combination of such paradigms. 

Cleversafe: revolution in storage

By Ilia Volvovski, Jason Resch
Illustrated by Alexander Zhelonkin

Published: #9 Spring-2012



The amount of data stored in the world is growing exponentially. This year, 2012, has been called the year of Big Data; companies and individuals generate enormous amounts of data in the form of pictures, movies, scanned documents; X-rays and other medical images; results of scientific experiments in physics, climatology, and biology; surveillance data from public places; financial information. It is estimated that the amount of digital information is doubling every 18 months, with 95% of this coming from unstructured data, with the remaining 5% being driven by traditional structured data (databases).

The trend of unstructured data growth is expected to far outpace the growth of structured data well into the future. The problem facing us today is: how can we keep pace with these exponentially increasing demands to store this information reliably, securely, cheaply, and easy to find and access? Storage: RAID is not the solution. The simplest storage solutions are based on a single storage device, such

as a hard drive, SSD, DVD, thumb drive, etc. These devices are, of course, prone to failure, and when they fail often all or a portion of the data they kept is forever lost. The most trivial way to overcome failures of storage devices is to replicate the data; that is to store multiple copies so that if one device fails, the data can be restored. This, however, is expensive. For a single backup copy, one has to use twice as many devices, for two backup copies, three times as many memory devices are needed. For this reason, a technique called RAID was created. Developed in the 1980s, RAID (Redundant Array of Independent Disks) uses forward error correction techniques to enable the recovery of up to one disk failure (for RAID 5) or two disk failures (RAID 6) out of the total set of disk drives. For example, a RAID 5 array could consist of four drives, and could tolerate the loss of

any one disk drive from that set of 4. This provides similar reliability to making one backup copy (as you can now tolerate one failure) but it is much more efficient: its overhead in this case is 33%. However, as disk sizes have grown in capacity, several problems have emerged. In the early 90s, when RAID was in its infancy, it took approximately one minute to read all content on the entire drive (which was only 40 MB). This meant that a RAID array could be rapidly repaired in the event of a disk failure. However, by 2010, drives had grown about 50,000 times in capacity, while disk read speeds had only increased 150 times. The result is today, it can take between 6 to 8 hours to read an entire disk, and it may take days to rebuild a failed drive in an active RAID array. Since the window of vulnerability increased by a factor of about 500, the probability of

In the early 90s, when RAID was in its infancy, it took approximately one minute to read all content on the entire drive (which was only 40 MB). This meant that a RAID array could be rapidly repaired in the event of a disk failure

a secondary disk failure happening in that window has also increased by a factor of 500. The net result is that using RAID 5 today, the chance of data loss is 500 times greater than it was in the early 90s. This was the primary motivation to move to RAID 6 systems, which can tolerate 2 simultaneous disk failures. However, RAID 6 has turned out not the panacea it was thought to be. There are two common ways a disk drive can fail: it can fail totally, or it can fail silently during read back from a certain location (called a sector) on the disk. An operation succeeds but the result is different from the original written data. Silent sector errors used to be very rare, they occurred once for every 10 – 100 TB read. When disks were in the MB to GB sized-range, they were practically non-existent. However, with disks now 2-3 TB in size and RAID arrays up to 50 TB now common, the chance of a sector error occurring during a rebuild is significant, it can be over 50%. The impact to RAID 6 is severe; approximately half the time, RAID 6 is unable to recover from a double disk failure. Sector errors, in combination with the large drives of today make RAID 6 only marginally more reliable than RAID 5. All this factors contributed to abandoning pure RAID 5 systems from enterprise lexicon and the time is coming when RAID 6 should expect the same destiny. The approach the industry has taken to make up for RAID's deficiencies is replication. It is considered standard now to have at least three copies when storing valuable data. In this case, replication not of single disks, but of whole arrays or sites.

When dealing with small systems, replication is not too costly. However, when dealing with a PB sized system, for which disk drives constitute 70-90% of the cost, and a single copy costs millions of dollars, replicating this system will cost many millions more. The challenge is to create more reliable and available storage that is cheaper than the replication based option.

Access: make it simpler

Another aspect of the storage system is an access mechanism it provides. Typically storage systems provide block level interface. File systems are built on top to organize blocks of data into logical structure. Network Attached Storage (NAS) are built on top of file systems to provide remote access to the stored files. However file names and directory hierarchies were created to assist humans in organizing data in a manageable form. They were designed to allow concurrent access to relatively small groups of data files shared among a few users. Using NAS storage for unstructured data leads to higher price – software doesn't need hierarchies; directory structure is not natural for automation – they

How can we keep pace with these exponentially increasing demands to store this information reliably, securely, cheaply, and easy to find and access?

are not computer programs friends – unique long keys and associations are. Relational DBs perfectly serve this purpose, but they do not scale too well for hundreds of billions of objects and distributed processing, thus causing the proliferation of non-SQL DB in recent times.

A new scalable approach is object based, when content and metadata (data context) are stored together. It usually utilizes flat virtually unlimited namespace and is ideal for storing billions of relatively static objects.

Dispersed Storage

Dispersed Storage is a modern approach to achieve scalable, secure, reliable and efficient data storage. While based on math that has been known since the 1960s (Reed-Solomon codes), it is only in the past decade that CPU processing rates and network throughput has made such techniques viable alternatives to local storage methods based on RAID or replication. At this point dispersed storage may be the solution that can cope with continually growing storage requirements.

Dispersed storage is based on Information Dispersal Algorithm (IDA) that is based on math known since 1960 (Reed-Solomon erasure codes). It relies on a simple matrix algebraic fact. If you have N values, they could be encoded with a NxM matrix (where M is equal or greater than N; M is called IDA width and N is called IDA threshold) such that any NxN sub-matrix could be used to solve a linear equation and thus to reconstruct the original data. Each stored unit that would be used for reconstructing original data is called slice and is 1/N of the size of the original data. Since we need to store M slices the storage overhead in this case is M/N, a value greater than one. The same overhead may lead to a vastly different reliability. For example both M=3, N=2 and M=30 and N=20 (M is called IDA width, and N is called IDA threshold)

have the same overhead of 50%, but the second system is much more reliable – up to ten nodes could be lost simultaneously without losing the original data. It should be noted that number of combinations that lead to loss is growing but not as fast as the probability of simultaneous losses decreases. In case of 20 of 30 the number of combination is around 30 million (~3*10⁷), a probability of 10 disk losses is p10 (where p is probability of a single disk loss and is considered less than 10⁻²). This results in value less than 10⁻²⁰.

Dispersed Storage eliminates the need for replication, by making a trade-off in CPU. RAID 6 computes redundant information so two disk failures can be tolerated. Dispersed Storage enables an arbitrary amount of redundant information to be generated, and thus can support an arbitrarily high number of simultaneous drive failures. The relationship between CPU power and redundancy to be generated in IDA is linear. Each additional failure that can be tolerated increases the system reliability by a factor of about 100. Thus a system that tolerates 6 failures will be 1004 times more reliable than RAID 6, yet it only requires 3 times as much processing. As Moore's law continues, the cost of this processing becomes increasingly marginal, especially in comparison to the millions saved by avoiding replication.

Cleversafe

Having a theoretical foundation is the first important step for a commercially viable solution but in no way it is the last.

Cleversafe is the first company to provide commercial storage based on IDA. Cleversafe was founded in 2005 and was steadily growing ever since. Our first software and hardware release was announced in February 2008. It only supported SCSI interface to dispersed storage (dsNet). It means that a single data container would expose dispersed data as a block device. It makes possible to create and

mount a file system on top of dsNet. No integration is required for the existing applications relying on file system interface. As seamless as this approach is its potential is limited to up to hundreds of TB installations. It was not the scale we anticipated.

In 2009 Cleversafe introduced its first object based storage. Objects of any size could be stored on dsNet and they are uniquely identified by a global ID. It is the most efficient and scalable way of storing data on dsNet, however it moves the responsibility of maintaining IDs to a dsNet client. In many cases it is not a limitation as host systems already provide capability of storing object related information. For example, Shutterfly – the largest online photo-sharing system in the world and the current Cleversafe largest client – has a DB

When dealing with a PB sized system, for which disk drives constitute 70-90% of the cost, and a single copy costs millions of dollars, replicating this system will cost many millions more

that stores information related to each picture such as an owner, date, a title, comments and photo related metadata. A digitized content of each picture is saved in Cleversafe dsNet as an object and its unique ID is stored in DB along with other information. A lot of companies realize that this is the most sensible solution for their needs, although their current systems may use traditional NAS to store the same objects as files.

Cleversafe current focus along with simple object (SO) storage is to provide industry standard interfaces to dsNet such as S3. It would require more than storing a simple ID. The other development effort is to integrate dsNet with existing storage systems to provide an alternative and efficient back-end (such as MS SharePoint, IRODS). Cleversafe is also focused

on building next generation systems that would grow upon the existing technology to provide unlimited indexing scalability with properties that are inherent to dispersed storage: reliability, availability and unlimited scalability.

Design principles

Our major guideline in designing dsNet was to provide unmatched system reliability and availability. These are natural extensions of IDA based dispersed storage as described earlier. Other goals were:

- *Infinite scalability. This means that the system should not expose any single point of failure; no central authority should reside in data I/O path; all system functions should be scaled with the number of storage nodes; no*

artificial limits of storage capacity within single container (which is called Vault in Cleversafe lingo); system throughput should grow with system capacity. In other words, as dsNet capacity grows so does its ability to execute its internal functions. It is a very important distinction as most if not all of the existing large storage systems rely on a central coordinator for routing and other functions.

- *Data integrity and consistency. dsNet storage guarantees that if an application had successfully written data it will be able to recover written data in its original form in presence of failing disks, unreliable networks, failing applications.*
- *Threshold security. Unless an intruder gains access to at least threshold number (N value in IDA) of nodes they could not decipher any information they obtained.*

However if a threshold number of data is obtained the original data is fully accessible, in other words this mechanism does not replace cryptography.

Geo-dispersal

When a system can tolerate many simultaneous drive failures, the most likely route to data loss is not simultaneous independent disk failures, but rather correlated failures such as a fire, a flood, or other type of disaster. Thus to fully realize the reliability provided by IDA, it is necessary to make the failures as independent as possible. This is achieved by using entirely separate computers (nodes) in the place that independent drives held in RAID. Some have called this approach RAIN (Redundant Array of Independent Nodes). The advantage of this is greater availability and a reduction in correlated failures. If one node goes offline or is destroyed, corrupted, hacked or otherwise fails, it is tantamount to a single disk failure in a RAID system. Although the idea of geographical dispersal is not original (for example copies are typically stored in different physical locations), it is organic for dsNet as by its own nature requires multiplicity of storage nodes. Since nodes communicate over the network, they can be placed anywhere. Even across the globe. This deployment is something we call geo-dispersed storage. It takes the reliability to a higher level, as a storage system becomes resilient to natural disasters and physical attacks.

Data integrity and consistency

Due to distributed nature of dsNet data integrity is by no means a trivial task. Multiple nodes are involved in a single logical operation. While operation is in progress a participating node could go down, disks could fail or network

When dealing with a PB sized system, for which disk drives constitute 70-90% of the cost, and a single copy costs millions of dollars, replicating this system will cost many millions more

connectivity could be lost. The modification operations on dsNet (write/delete) are fully transactional and are implemented with three phase commit. First data is written on all nodes; second data is committed and becomes visible and lastly old data is discarded when sufficient number of nodes confirms that newly written slices are durable.

High performance

Cleversafe software is built to satisfy high performance requirements of existing applications. All operations implemented in concurrent manner; participating nodes read and write object's slices concurrently thus reducing latency and improving throughput. Since multiple slices are required to save/restore the original data they are written/read concurrently thus improving system throughput and naturally achieving load balance. System throughput for a SO vault could be improved by increasing number of access points and is limited to disk I/O speed of a single storage node and network throughput. As capacity grows the number of storage nodes increases and so does the overall performance.


Rebuilding

dsNet is tolerant to disk failures, device losses and temporary connection failures. In order to maintain high level of reliability

missing or outdated slices should be promptly rebuilt. In case of replication based storage it is quite simple: one has to recreate a full copy of existing data on a replacement node. For dispersed storage a process has to recreate missing slices from existing ones. This means that rebuild software has to read threshold number of slices in order to rebuild one.

Rebuild process consist of two parts: continuous scanning and discovery of missing or outdated slices and actual rebuilding. This is done by a background process active on each server, so when a storage system grows so does its capacity to rebuild slices. It should be noted that due to significant redundancy (usually 6-10) there is no high urgency to recreate slices. Even a one month window represents a very low risk of losing user data – orders of magnitude lower that of RAID 6 within 8 hours rebuild window. As the result the system could find the least busy hours to perform rebuilding more actively.

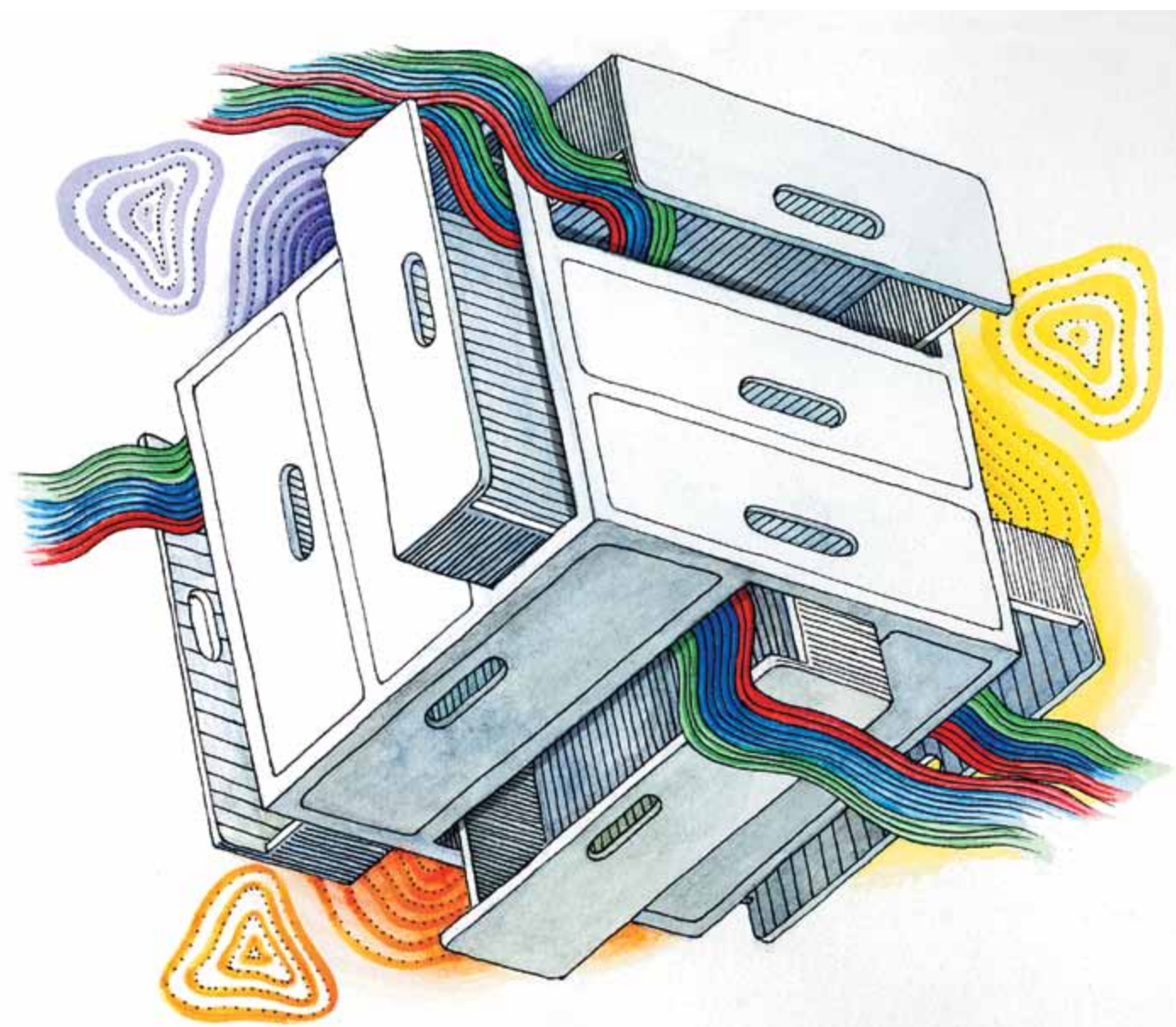
Security

Several mechanisms could be employed to provide secure access to stored data. dsNet supports username/password and PKI authentication. Different roles could be assigned to various groups of users to execute various types of tasks. Data security at rest is achieved by deploying All-Or-Nothing-Transformation (AONT) encryption methodology that guarantees threshold cryptographic security without imposing an external key management. It is not security in full sense of the word as anyone in possession of threshold number of slices should be able to recreate the original data. However it reduces requirements for disk protection as a single disk doesn't reveal any information about data it is part of. 

SciDB – new DBMS for large amounts of scientific data

By **Oleg Bartunov, Pavel Velikhov, Roman Simakov, Konstantin Knizhnik, Artem Smirnov**
Illustrated by **Alexander Zhelonkin**

Published: #5 Spring-2011



The SciDB Project resulted from the assessment of the current situations in large-scale scientific projects by the leading scientists of various sciences, the representatives of commercial companies and DBMS developers (database management systems). These issues were discussed at a number of XLDB conferences. This Project is managed by the distinguished MIT Professor Michael Stonebreaker with his colleagues from the largest US universities. This project started in 2008. Apart from American scientists, it involves Russian scientists and developers from the Research and Development Institute of System Research of the Russian Academy of Sciences and the Sternberg Astronomical Institute of Moscow State University. Michael Stonebreaker is the author of the first relational DBMS Ingres, DBMS Postgres and a great number of other successful database implementations.

The main goal of the project is to develop DBMS for the needs of large-scale scientific and industrial projects as soon as possible, which require the analysis of very large data volumes (hundreds and thousands of petabytes) scaled for thousands of servers.

The SciDB system has a number of fundamental differences from the existing DBMS. SciDB is developed as a system for storage and analysis of raw and derivative scientific data. Some basic functions of traditional databases are not supported by SciDB, allowing the system to perform more efficient processing of analytical queries. For instance, as no initial data is actually updated, SciDB does not provide effective support of large amounts of transactions, which, in turn, allows avoiding considerable overhead costs. Finally, SciDB is a project with an open source code and a free license for use, which meets the requirements of the majority of customers. The open source code provides cutting the customers' costs for large-scale system implementations, and the open development process ensures high quality of engineering solutions.

Besides this, the transparency of DBMS provides processing independence and a possibility of data exchange among various research teams.

In addition to the traditional functions of the database management systems, SciDB employs new mechanisms for operations with data, which have been specially developed for scientific data analysis. The SciDB data model consists of multidimensional nested arrays. Since SciDB will store the data obtained from the scientific instruments, SciDB supports the measurement error at the level of the data model and query language. Finally, SciDB is primarily developed for operations with a wide range of computer systems, varying from a portable PC to large-scale clusters and supercomputers. Therefore, scientists will be able to work with the data in the same environment, for example, developing analytical algorithms at personal computers and using small-scale data samples, while the final unchanged queries will be launched on highly-efficient clusters. Also, SciDB is integrated with popular computational software packages, such as R, Matlab, etc, enabling scientists to use ready data processing algorithms when switching over to SciDB.

Main Specifications of Developed DBMS:

- *Raw data storage; processing occurs in DBMS by means of user-defined procedures to ensure version control and data change history – full-scale support of the complete operation cycle for scientific data.*
- *Description model for scientific data is a multidimensional nested array with ragged edges.*
- *Declarative query language for operations with arrays.*
- *Vertical (attribute) data storage for compression and reduction of input-output operations.*
- *Data safety due to the replication of some data at different nodes of the system.*
- *DBMS scalability from the notebook up to thousands of servers which can store tens of petabytes.*
- *Expandability with user-defined datatypes and functions.*
- *Refusal from supporting transactions, which are unnecessary for scientific data (WORM*

– Write Once Read Many) and significantly complicate DBMS architecture, thus increasing their maintenance costs. Specifically, this will help avoid logging and a large number of locks available in traditional DBMS.

- *Free license (GPL 3.0).*

SciDB is highly scalable parallel DBMS with «shared nothing» architecture – meaning that , each SciDB node operates only with local data and memory. Raw data processed by the cluster is distributed by the system as overlapped array chunks; the chunk and overlap sizes are customizable. The cluster is managed by the supervisor for each query, so the system may contain several supervisors, depending on the system workload. Each supervisor compiles queries into the plans and sends them out to the worker nodes performing queries on the basis of their local data. If any additional data is required at the worker nodes, Scatter/Gather operator is included into query plans for sending out additional data. Let's study an example from astronomy, where images from the charged-coupled device (CCD) matrix are stored in SciDB. Each node stores a particular subset of the data, and the data are represented as an array of tuples on the logical level, <x, y, l, err> in our example (x, y – are the coordinates of points at CCD, where l is the value of a signal, err – is a measurement error). On the physical level, the data arrays of various attribute are stored separately. Such model was dubbed vertical and is often used in contemporary DBMS. The main advantages of such model include better compression characteristics of the data, as normally the same-type data is compressed better; during queries, only required attributes are selected for calculating the query. The physical level also implies chunk storage of the arrays of such attributes, or large pages also including overlaps. Therefore, a chunk with an overlap is considered a unit of data processing in SciDB. It is well-know that scientific queries over arrays of data often require local calculations, or calculations in a small area surrounding each point. Image smoothing filters are often used in image analysis. SciDB supports queries,

including neighborhood point analysis, by means of chunks with overlaps. If the number of overlapping data is sufficient for a query, SciDB performs parallel queries on all nodes. If overlapping is not sufficient, Scatter/Gather is included in this plan, which, in its turn, increases overlap.

The data storage model in SciDB implies replication to ensure fault tolerance of the system. Each chunk of arrays stored in SciDB is replicated at other nodes; during SciDB operation, the supervisor checks all working nodes involved to detect any node failures in due time. In this case, backup data stored at other nodes are to be used for query processing.

When developing SciDB, it was noted that scientific data is hardly changeable and access to all data versions is necessary. However, unlike commercial DBMS, the support of a large transaction flow is not required. That is why, when developing the system, one could avoid using a large number of synchronizing mechanisms available in traditional DBMS. This considerably facilitates the system operation, especially in the parallel data access mode, in which traditional systems have to use distributed lock managers. The declarative query language for arrays provides transparent access to multidimensional data - ranging from small datasets to petabytes. SciDB uses AQL (Array Query Language), which in many aspects is similar to SQL. AQL has the same SELECT-FROM-WHERE template, but operating with arrays instead of relational tables. The AQL query language enables to formulate queries analyzing point neighborhoods. Overlapped chunk data distribution ensures parallel performance of such queries without data movement between worker nodes. One of major AQL operators is REGRID, which is very much similar to the popular MapReduce. REGRID creates a new data array based on the original version, using two functions – domain and aggregate. The domain function selects the array subset for calculations, and the aggregate function calculates the value of the output array element. When using REGRID, it is much easier to select

points from their neighborhoods, while the chunk storage model with overlaps allows calculations of the output array at all nodes in parallel, without any communications. Let us consider an example of using the REGRID operator for smoothing the master data by means of a Gaussian filter.

```
SELECT l FROM CCD AS C
REGRID (
SELECT l FROM CCD AS C1 WHERE C1.i
BETWEEN C.i-20 AND C.i+20
AND C1.j BETWEEN C.j-20 and C.j+20,
SUM( C1.l * a*e^(( -i-b1)^2/(2*c1^2) + (-j-
b2)^2/(2*c2^2))))
```

In this case, the first parameter in the REGRID operator selects a 40x40 neighborhood of points around the input array point, while the second operator calculates the Gaussian function on the basis of the selected points.

Such queries are impossible in relational databases. Even simple data sampling, when the user needs to choose a data cube performed by the domain function, is already a complicated task for relational databases. For this purpose, relational bases require a multidimensional index or they need to go through all table data, whereas it is an embedded function in SciDB.

Such queries are principally possible in the MapReduce system – one may use a function map to select the required point neighborhoods and the function of reduction used for calculating aggregates. However, this complicates the user's work and degrades efficiency: it can be quite difficult to prepare a function map equivalent to the regrid domain function and select chunks with overlaps, plus the map scatters the data coinciding by their keys and nodes for simultaneous convergence. Regrid uses initial data decomposition and does not require re-distribution by nodes in most cases. Therefore, a regrid is a more convenient and at the same time is a more effective MapReduce version for operations with scientific data.

Full-Scale Support of the Complete Operation Cycle for Scientific Data


As it was mentioned above, due to a

number of disadvantages of the existing DBMS, the majority of scientific projects aimed at the analysis of large data volumes perform master data processing and analysis outside the database management system. SciDB can solve this problem by providing highly-efficient and user friendly master data warehousing and a large set of tools for data processing and analysis. The versioned array storage and the history of all data transformations enables SciDB users to receive accurate information about data versions and all computations made with the raw data. This allows for effective error elimination in data processing algorithms, tracking master data processing in case of any suspect results and replicate calculations with raw data. In addition, SciDB operates without any restrictions, both at supercomputer clusters and personal computers, which enables scientists to operate in the same environment with his or her data. The SciDB system is also capable of tracking data lineage.

When processing the raw data and its derived objects, SciDB memorizes queries used for obtaining specific results. By a DBMS user's request, the system may replay the process of obtaining results from the raw data or provide database sampling required for obtaining such results.

After processing the raw data, SciDB allows to share the obtained results, to provide samples and to perform analytical queries for a wide range of colleagues, complying with the sampling policy both for the data and the obtained results.

Therefore, SciDB supports the full cycle of data processing and analysis, from storing the raw data to analyzing the obtained results. In this case, all results obtained through DBMS can be replicated and reproduced by the user of the system.

Up to now, version 0.75 has been released, still containing a lot of restrictions. However the tests using realscientific data and typical queries have already proven the best efficiency, as compared with relational DBMS. The full-scale version 1.0 will be released in May 2011. 

On a motorway with two lanes

Editorial interview

Published: # 7, Autumn-2011

The main thing is to set the direction
The editor of our magazine Igor Levshin talked to Jean-Yves Berthou. Coordinator of the European Exascale Software Initiative (EESI) and is also director of IT development at EDF - the leading French energy company.



Igor Levshin: Jean-Yves, would you please share a little on the HPC development at EDF.

Jean-Yves Berthou: EDF have developed simulation tools for dozens of years. Areas range from power generation in nuclear and hydropower plants to simulation of electrical networks, the study of renewable energy technologies, security facilities and minimizing pollution (these studies must be conducted by EDF on a regular basis, according to the established plan). By the way, we have developed many applications in-house, have written millions of lines of code and remain the owners, although it is shared with others.

I. L.: Which of these applications are the most demanding in terms of resources consumed?

J.-Y. B.: CFD and material simulation.

I. L.: Prior to embarking on your EDF career you worked for CEA – France's National Nuclear Agency. Could you compare their approach to HPC?

J.-Y. B.: These are very similar areas facing lots of similar problems. Many projects are done jointly. The nuclear agency CEA traditionally performed lots of scientific-research work and many fundamental studies. EDF also has invested greatly in the research, but there are more practical calculations related to the efficiency of production and transmission of

energy.

I. L.: How do you manage to balance the interests of science and business in the EESI?

J.-Y. B.: In Europe the situation with the HPC has been changing very fast

over the last few years, largely thanks to education in this area. It concerns both the attitude towards HPC in general and specific HPC projects. Today France has many areas of the economy where large corporations are investing in research, including academic studies. Many researches are conducted, of course, with the use of HPC. In addition to EDF, Total (oil and gas industry) and Airbus (aircraft industry) collaborate with research institutes, funding not only research which is of particular interest to them, but also academic research. The three companies all have their own HPC peta-scale facilities and projects.

I. L.: How did you manage to meet the needs of European vendors and multinational manufacturers of supercomputers alike?

J.-Y. B.: We are doing our best to assist the development of European IT companies; we collaborate with them, exchange information, and assist in the research of capabilities in the field of parallel computing. We attempt to create a European HPC ecosystem. But we cannot afford to rely only on European companies because sometimes their capacity is not adequate. And important studies must be carried out without delay. We make a tender. But when it comes to bidding, then we provide very fair competition without any protectionism. Overseas companies such as, say, IBM or HP, participate in tenders on equal terms. We have many examples like this.

I. L.: How do EESI and International Exascale Software Project correlate?

J.-Y. B.: We are deeply involved into IESP, and we are working to ensure that Europe's contribution to HPC increases. We are committed to cooperation on a truly international basis. We are well in touch with the leaders of IESP, Jack Dongarra and Pete Beckman, sharing plans and together we participate in the organization of meetings to protect the interests of our initiatives at the national level.

I. L.: Right now, we are looking forward to another international event – the

Supercomputer Conference and the ISC 2011 exhibition in Hamburg.

I. L.: Do you have similar views on the exa future?

J.-Y. B.: Yes, I think so. We have synchronized our programs to match the pace and expectations of IESP and no dissonance has so far occurred.

I. L.: What is your attitude towards the hype of GPUs - do you think they will become part of the HPC mainstream soon?

J.-Y. B.: No, I would not rush to answer this question. And in general, it makes sense not to talk such abstract things, as it depends on the industries and typical tasks. For example, in the oil and gas industry, GPUs are used very actively (in companies such as Total or Schlumberger) and it makes sense, because graphics accelerators work really well for solving typical geophysical problems and even today they need systems with a total performance measured in petaflops. Graphics accelerators are the best option for them. But for many tasks that are fulfilled in EDF, GPUs do not serve any good purpose while versatile CPUs often work better.

In general, it is clear that future systems will be of two types: one with architecture, which has already become a supercomputer classic and another more specialized equipped with accelerators (GPUs or even based on FPGAs). It will look like a motorway with two lanes. And we will see later, which lane will have more traffic and how easily you will manage to switch lanes.

I. L.: Do you think that the FPGAs can be very important in the HPC?

J.-Y. B.: I don't know. I only know that Russia has very impressive achievements in this field - interesting research with practical results. Perhaps this is the only country where so much attention has been paid to FPGAs. Let's see what happens next.

I. L.: What do you think about the possible contribution of Russia into this industry?

J.-Y. B.: At this point, due to Russia's dynamic development and growth, I

am not sure. I had a conversation with Vladimir Voevodin. He is involved in our project and his activities have left a very good impression. It seems that what is being done in Russia for education in the field of HPC has been done properly.

I. L.: Do you think whether it is needed – as Thomas Sterling put it – a revolution in software to achieve the exascale? Or will simple evolution do?


J.-Y. B.: Such disputes have been going for a long time, I know about them and frankly I am not very worried about any of this. It all depends on the application. When we consider applications, it is definitely a revolution. The code must be rewritten to adjust it to parallel computing, we need new languages, compilers and brand new infrastructure software. For those who have many applications running using old code, it's a large problem. If we talk about the supercomputer itself and system software, then what it would be like – well regarding evolution or revolution? It's hard to say.

I. L.: What do you think about alternative technologies: quantum computing, optics and others? Are these areas outside the scope of your interests and EESI?

J.-Y. B.: We maintain relationships with major hardware manufacturers, so we follow the course of various research work done. Optical elements within computer systems will be key components for example - but I do not think that fundamental changes in the systems will arrive before the year 2015.

I. L.: What has EESI undertaken in the field of education? It is an important activity - do you agree?

J.-Y. B.: Yes, of course. But we are not looking after the learning itself so much, but rather creating the conditions for the proper training. Working groups are conductors of ideas.

Our main role is to ensure international cooperation and synchronization of efforts. The main thing is to set the right direction. 

Siberian Federal University (SFU)

Contributor to the National Supercomputer Platform, Member of the Supercomputing Consortium of Russian universities

www.sfu-kras.ru

Published: #10 Summer-2012



Since its founding, one of SFU's development priorities has been to create a powerful high-performance computing center to provide modern tools of scientific and engineering calculations. In 2007 the university installed a supercomputer, which was put on the TOP500 list. Currently, the supercomputer complex of SFU is on the well-known list of Top50 Russian supercomputers. In 2010 SFU led the rating of traditional and national research universities in terms of innovation and commercialization of products.

High-Performance Computing Center

<http://clustersfu-kras.ru> (<http://cluster.sfu-kras.ru/?lang=en>)

Director – PhD in Technical Sciences, D. A. Kuzmin.

This Center serves SFU's research teams, interacts with knowledge-based enterprises of Siberia, Branches of

Russian Academy of Sciences, as well as Russian and European universities. The Center consists of three high-performance systems that integrate 280 computing servers, SAN network of 8 Gbit/s, InfiniBand 20 Gbps /s. The hybrid component of the

complex is platform NextIO Vcore Express with 4 GPU Tesla M2090. They have installed software packages that allow to conduct resource-intensive calculations in the various fields of science. The installed software includes ANSYS, Matlab, GAMESS-US, CFOUR, MRCC, NWChem, ORCA, AMBER 11, and SigmaFlow. All the software products have academic license or are open-code products.

In order to provide remote access to high-performance resources, SFU has fiber optic communication channels providing data transfer speeds of up to 10 Gbit/s and an access channel to public Internet networks at the rate of 200 Mbit/s.

The Center has actively collaborated with such companies as IBM, Intel, NVIDIA, AMD, Cisco, VMware, and Novell. Its researches and specializes in the management of high-performance systems. The center has been implementing the project «High-performance computing as a service». This project aims to put in place a versatile data-processing infrastructure embedding a number of services that address specific applied tasks. User access to distributed resources through this Center is enabled through one web portal. They also have implemented a comprehensive management system allowing administrators to efficiently manage computing resources, roll out new specialized software, monitor resources and user tasks.

The Center has implemented a 3D-rendering service (www.rendermama.com) focusing on a wide range of users working with 3D graphics including designers, animators and architects. The system supports the loading of the projects in batches 3DMax, Maya, and Cinema4D. Rendering of projects is made in automatic mode on distributed supercomputing resource center using software MentalRay and V-Ray.

SFU High-Performance Computing Center is a base for training masters and bachelors in the field of supercomputing technology (SCT)

Fundamental and special courses in SCT are read by leading experts of the Institute of Mathematics (IM) and the Institute of Space and Information Technology (CITI) of the SFU. They also have an accredited Master's program on «High-performance computing systems». They also have put in place the professional training program «Parallel Programming». SFU research teams have carried out fundamental and applied researches in the field of supercomputing technologies including.

Instrumental and language support for architecture-independent parallel programming

A research team led by Professor A. I. Legalov has solved the problem of creating architecture-independent parallel software. The proposed approach is based on functional-threading parallel programming paradigm, which enables the writing, debugging, and verification of programs without any reference to specific computing resources. The team has used their own research tool, a programming language «Pythagoras». Currently, they study the possibility to the use this language

as a high-level add-in for FPGA design, as well as for parallel programs running on multicore processors.

Technology and tools of architecture-independent design of high-performance single-chip systems

One of the main activities of the research team of specialized scientific and educational Microprocessor Systems Laboratory led by Professor O. V. Nepomnyashiy, is the creation of new technology for the design of reconfigurable and dynamically reconfigurable systems-on-a-chip (SoC and DRSoC). Methods and tools of high-level design and verification of high-performance systems on a chip that have been developed in the course of scientific research, facilitate effective work on SoC projects, creating ready-made solutions, software and hardware of the system

and topology of single-chip systems which, on one hand, do not depend on the back-end implementation and allow to maximize the level of abstractness of functional algorithms from the architecture of the target ASIC, and, on the other hand, allow to draw as close as possible to the representation in the hardware description languages and are able to be adequately applied to the physical design of SoC.

With this approach, developers do not operate either at the level of ready-made hardware platforms, or at the level of previously developed and verified IP blocks, rather, at the level of the principles of system organization of computing process with an effective transfer of the obtained model to the target ASIC. The research results are applied in the development of systems for space exploration, including energy management systems and executive automation systems in spacecraft.

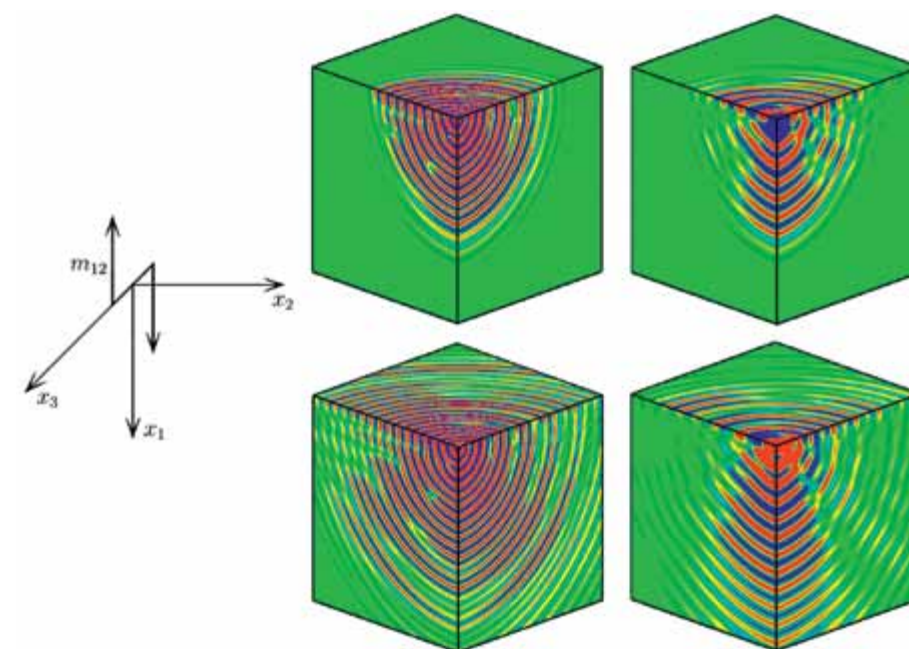


Fig. 1. Load distribution and the surface level of the angular velocity for the non-resonant (left) and resonant (right) frequencies at different times

Computer simulations of large-scale algebraic systems, computational group theory (CGT)

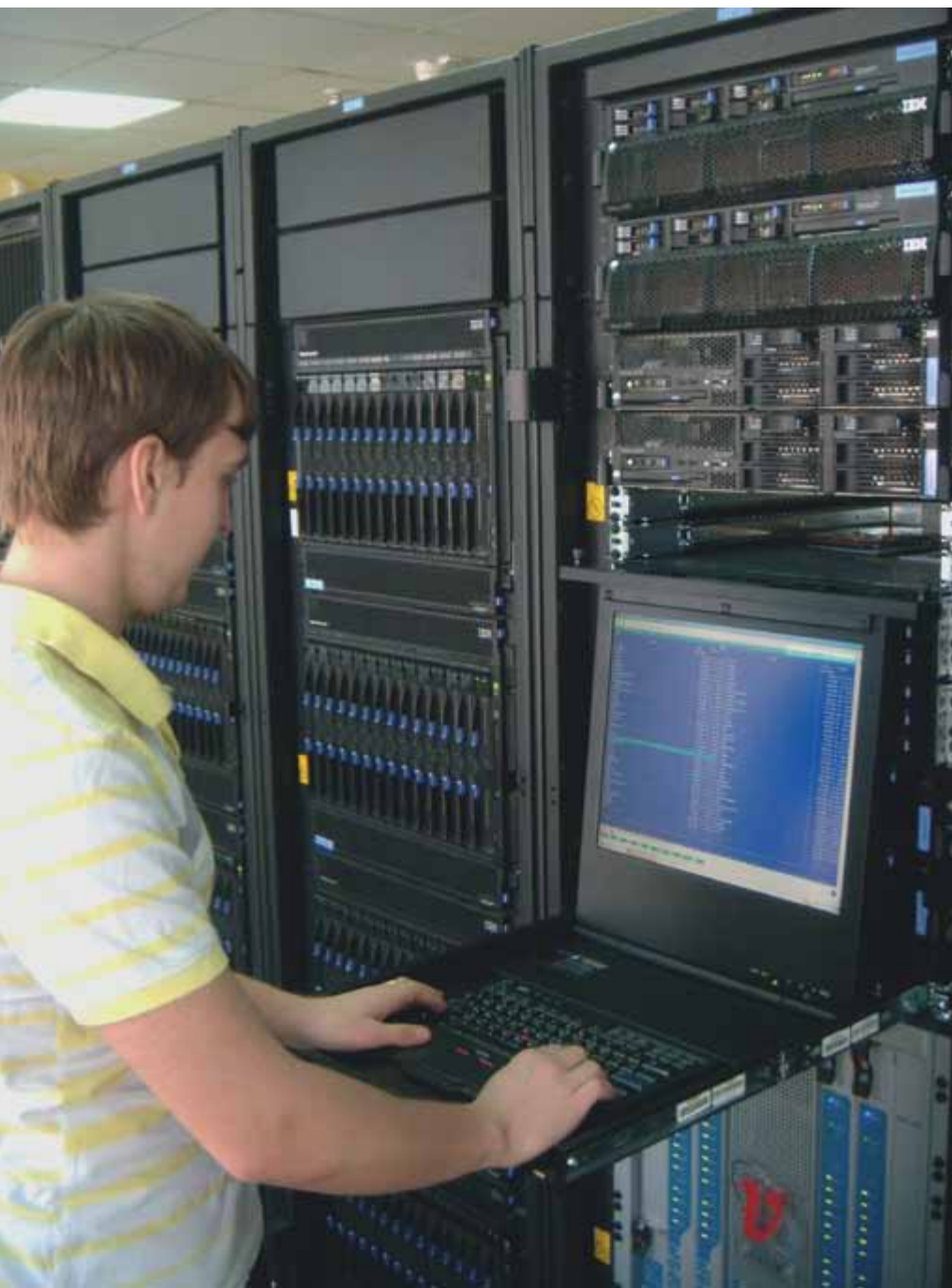
Studies conducted by research team led by DSc A. K. Shlepkin and DSc A. A. Kuznetsova include design, analysis of algorithms and data structures for the calculation of various

characteristics (most often – target values) of respective groups. This field is interesting due to investigation of groups that are most interesting from different perspectives and the data on which are not available via manual calculations. A set of algorithms and programs for solving the problems of group theory has been developed. As a result, the Charles Sims problem on

the structure of relations in the two-generated free Burnside groups of period five – Group B (2.5) has been solved by a SFU supercomputer, as well as the problem of the recognition by spectrum of special projective linear group over a field of dimension two of the seven elements – group L2 (7). The team also has studied the structure of a number of groups with different conditions of saturation.

Parallel computing in problems of continuum mechanics

The Institute of Mathematics at SFU has actively collaborated with the Institute of Computational Modelling with the Siberian branch of Russian Academy of Sciences. At the basic Department of Computing and Information Technology, which is headed by corresponding member of the RAS, V. V. Shaidurov, students and magistants are involved in several integration projects of the Siberian branch of the Russian Academy of Sciences for the solution of problems of continuum mechanics with the use of high performance computing. For instance, a group of MI SFU and ICM of the Siberian branch of the Russian Academy of Sciences managed by Professor V. M. Sadvskii develops and explores mathematical models of mechanics of deformable non-classical media. The models consider different tensions and compression resistance and structural heterogeneity of the environment (soil, rocks, coal graphites, polymers, and porous media). A set of applied parallel programs for solving tasks on the distribution of stress and strain waves in media with complex rheological properties on cluster-architecture supercomputers has been developed. Fig. 1 shows the results of calculations of the spatial problem of periodic action of concentrated moment at the



The development of methods and algorithms for analysis and interpretation of satellite imagery

In order to solve the problems of the preliminary and thematic processing of satellite imagery, a specialized MATLAB-based software package has been put in place along with a system for receiving, processing and storage of satellite imagery completely covering all the area of Krasnoyarsk Region.

They also have solved a number of problems related to segmentation of clouds on satellite data, the calculation of vegetation indices for agri-landscape zoning.

This work has been conducted by scientific team of CITI sponsored by the Ministry of Informatization of the Krasnoyarsk Region.

SigmaFlow – the solution of problems of hydrodynamics, heat and mass transfer, and combustion

Specialists from the Krasnoyarsk branch of the Institute of Thermophysics of the Siberian branch of RAS, The Department of Thermal Physics of the Siberian Federal University and LLC «TORINS» have developed a SigmaFlow program designed to be the solution for a wide range of problems in the field of hydrodynamics, heat-mass exchange and combustion.

SigmaFlow allows to create the geometry of the design object, prepare computation meshes, perform computation itself and analyze simulation results using graphical tools.

This program enables calculations on clusters running on Windows or Linux.

Numerical method embedded in this software is based on the finite volume method for unstructured meshes, which ensures a conservative nature


of algorithms and allows to simulate processes in geometrically complex objects. In order to approximate differential equations they use stable schemes of high-order accuracy. The correlation between velocity and pressure fields is enabled by the procedure of splitting. Systems of difference equations are solved by iterative method using multigrid methods. The team of «SigmaFlow» developers have acquired vast experience in calculating «real» applications. There is a training version of the program.

Simulation of mechanisms of chemical and adsorption processes

Scientific team including O. B. Hajiyev (of the G. G. Devyatikh Institute of Chemistry of High-Purity Substances at RAS), and A. I. Petrov (SFU) represents the new era of academic mobility as the work is carried out in remote mode, combining scientific resources of SFU, the RAS institutes, and major foreign universities. They have solved problems with focusing on simulation mechanisms of chemical and adsorption processes that have opened a vast avenue for improving our understanding of the nature of chemical processes, thus drawing us closer to answering questions about the origin of life taken as an open nonequilibrium chemical system.

The team has completed a full-scale simulation of adsorption processes on the ice surface that are important for atmospheric chemistry and simulation of prebiotic period in Earth's evolution.

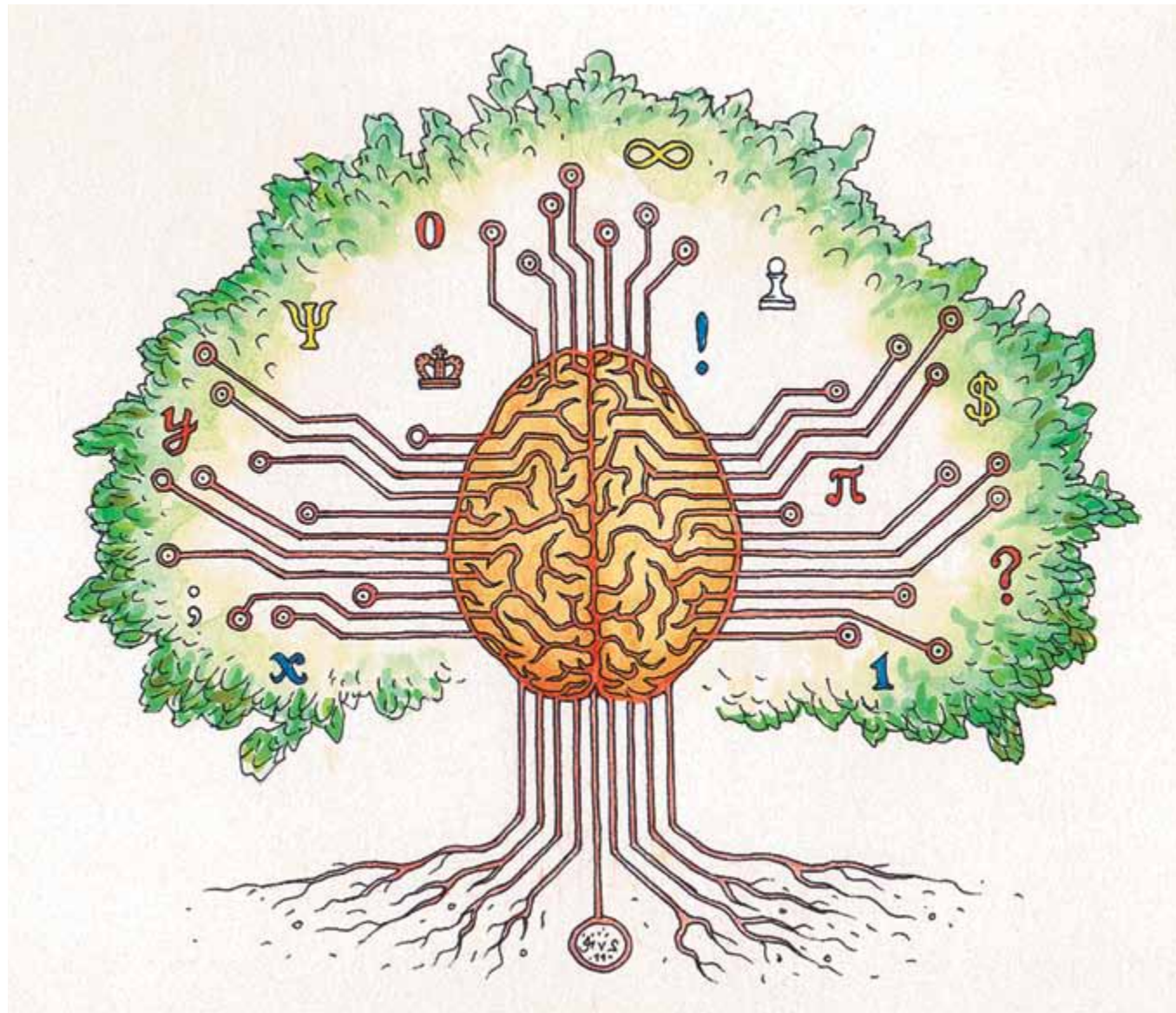
They have set a computing record for quantum-chemical study of reaction mechanism of $2NO + O_2 \rightarrow 2NO_2$.

They also found out the mechanism and the cause of stereoselectivity of the catalytic reaction of condensation in a system consisting of 100 atoms involving L-proline derivative. 

Computer numerical mathematical model and computer analogy method

By V.V. Aristov, A.V. Stroganov

Published: #8 Winter-2011-2012



Today it would seem impossible to imagine a theoretical science without the use of powerful computers and computing clusters. Though, of course, analytical mathematics tend to manage without these external devices – looking down on such «computing machines».

But the possibilities of “pure” mathematics in the field of computing are very limited. For example, most non-linear differential equations (which play an important role in applications) have no explicit solution. If we know the explicit analytical solution then we have the essentially complete information about the behavior of sought function, because one can see the solution as a whole, with its qualitative features, and asymptotics etc. – which is extremely important e.g. for the physical understanding of the described phenomena. On the other hand, numerical methods allow us to solve very complex problems, but obtaining such solutions involve a huge effort – one has to design software, debug it, conceptualize a series of calculations received, etc. People today are almost handcuffed to computers! We now rely too much on technology

and sometimes we cannot think independently which is unpleasant for mathematicians. Thus, we can identify the gap between “pure” and computational mathematics, which has existed for many years (a situation which is getting worse). But is it possible to try to lessen (if not eliminate) the gap, if to direct our attention to the formal features of the computer performance in order to extract the analytical result from out there? Until now, the theoretical thought has mainly tried to create logic circuits or languages for computing devices with their further

inevitable embodiment in the material (and alienated from the human being) image. As far back as the 1930’s basic logical prototypes of modern computers were found – in particular, well-known Turing and Post machines which for the time could be classified as computers. These logical models served as the theoretical base for the future computers. In recent decades, theorists have been widely discussing models of new computers, for example, quantum ones, and gradually, even at the simplest level, the elements of such computing devices begin to get “materialized”. Symbolic computations (or computer algebra) is widely known, they are embodied in MATLAB and other state-of-the-

art systems – it seems that analytical operations here are primarily “in the same direction”, i.e. from man to machine. They do not make any sense without a computer. However, one can try to change the vector and thus at least, partially balance this one-sided trend. Let’s try to build another – a “numerical computer model”, which would be based on the representation of numbers in a computing device and studying of their behavior during solving problems. For clarification, let’s consider how difference schemes (bringing the differential equations at the discrete level) work. They are actually a universal tool for solving nonlinear problems. In difference scheme the

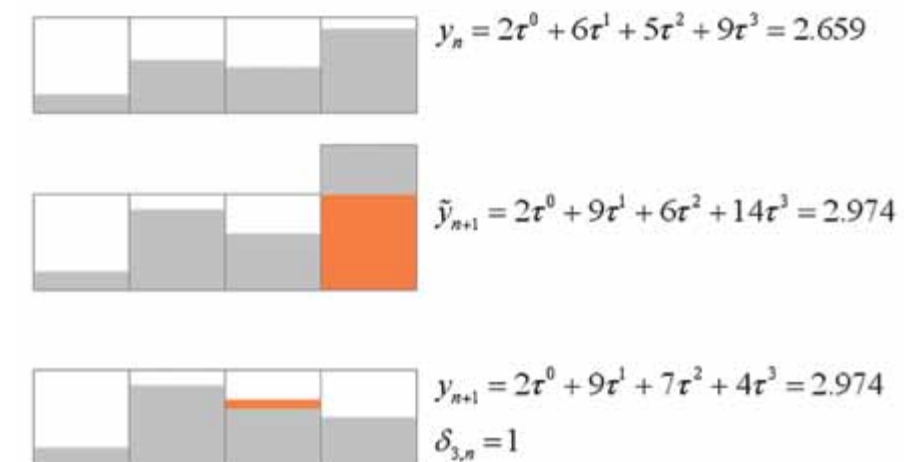


Fig. 1. An application of carry procedure with $\tau = 0.1$

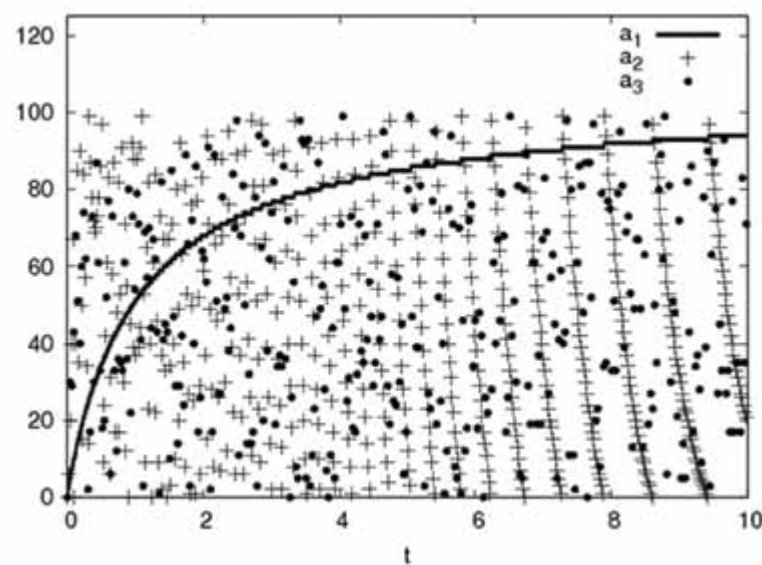


Fig. 2. A typical behavior of coefficients is the solution is represented as fragment of series in terms of powers of τ . Values of a_1, a_2, a_3 as coefficients with τ^1, τ^2, τ^3 respectively on Fig. 1

function value is expressed in terms of its value at the previous layer (of the independent variable), i.e. the recursive formula is used. To calculate the values of the function for large values of the independent variable, one has to calculate the recursive formula on a large number of layers - and that it is impossible to do without a computer. But is it possible to try, excluding the intermediate layers of the independent variable, to move to an explicit formula connecting the initial and final layers? It is interesting to look at the computer as at some black box into which we enter difference scheme, the original data, and the output is the number that appears in the result of a long multi-step process. It is clear how such a black box works at the algorithmic level.

Let's try to build a "numerical" model of computer with the help of general ideas of the representation of numbers in the computer. That is, let's try to describe its action during calculation of the difference scheme - not at the algorithmic level, but at the level of numbers and manipulations with them.

Let's introduce a step size of the independent variable, call it τ (this value can be considered small). Let's consider a theoretical computer model that would represent the numbers in a positional number system with base equal to $1/\tau$ (call it τ -computer). Then, at each step of calculation of the recursive formula the value of

the sought quantity will be stored in the form of a segment of series in terms of powers of τ . At each step, the values beyond the length of the series segment will be truncated. If these values do not exceed the error, that is allowed by difference formula (with some accuracy approximating the sought dependence), by order, then it can be argued that the representation of the solution in τ -computer also comes down to solving the original problem.

Similarly to operations in a classical computer, in the case of arithmetical overflow at some position, it is necessary to carry a part of the value to the other position, i.e. "from right to left". This procedure may be relatively easily formalized. Let at this step the solution in τ -computer be presented without overflows, and then an overflow in some position can occur only in the next step after applying the difference formula. Therefore, a value equal to the integral part of multiplication of this digit by τ will be carried to the neighboring left digit and a value equal to the remainder of the division of the digit by radix $1/\tau$ will remain. Thus the number stored in τ -computer will be redistributed among the memory

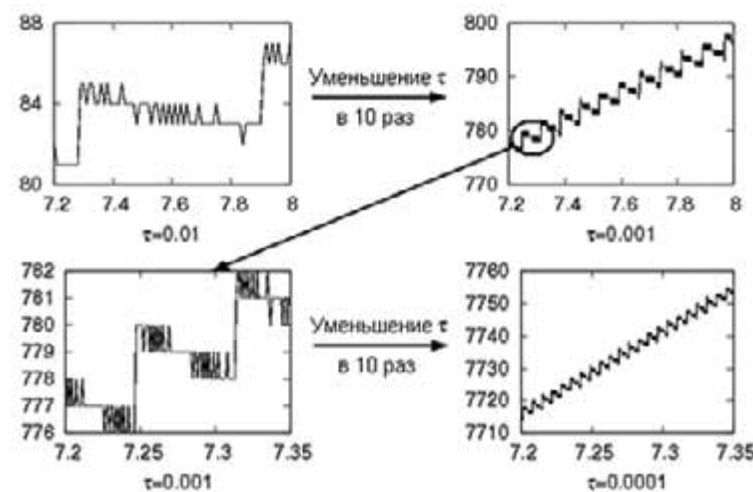


Fig. 3. Self-similarity features of carried values in case of depreciation of τ

cells, but will remain unchanged. This will help to maintain the desired error at each step.

Since the operation of taking the remainder of the division is used, the expressions for numbers in τ -computer number system generally correspond to the expression for the so-called pseudorandom-number generators, which are used to generate random numbers in the user programs and computer games. Some digits in the representation of solution in τ -computer inherit the stochastic behavior, which is confirmed by numerical experiments. However, not all digits behave randomly. The values of significant positions vary quite regularly, and the less significant ones - stochastically. Therefore, the number in τ -computer can be represented as a sum of deterministic and nondeterministic (stochastic) parts.

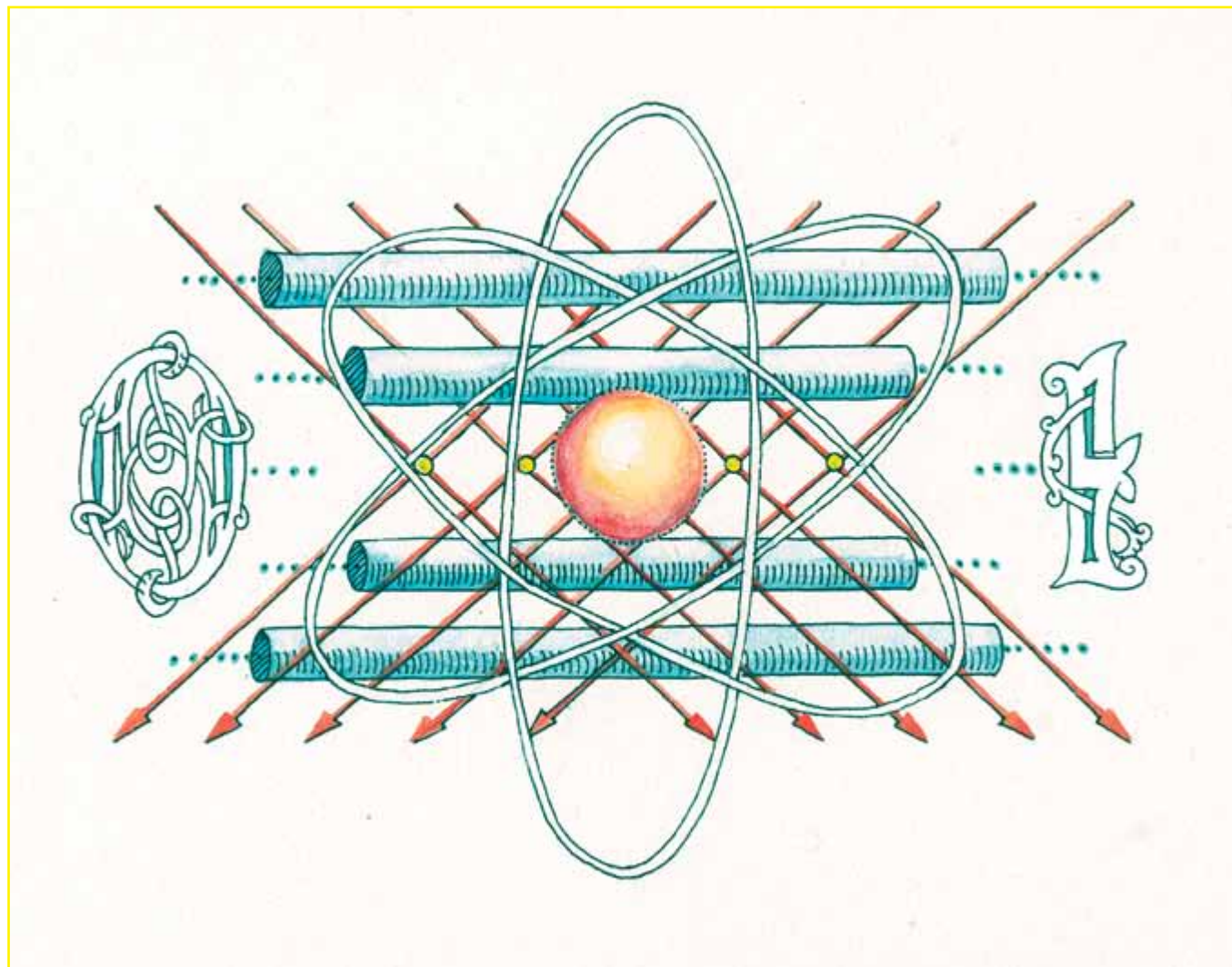
Therefore, using the methods of probability theory, one can predict how many steps it may take for the deterministic part of number representation to change. Thus all the intermediate steps in the recursive formula will be excluded. In practice, this results in an explicit inverse dependence of the independent variable on the value of the sought function. Direct analytical dependence can be obtained by the "idealization" of τ -computer, namely, in case of passage to the limit as τ tends to zero. Due to the novelty of the presented algorithm its application may be subject to considerable analytical work. Nevertheless, the algorithm has several advantages, in particular - it can be faster than conventional numerical methods; there is a possibility of efficient parallelizing, identification of the qualitative properties of solutions, obtaining simple asymptotic behaviors. But the main thing is that it helps (this requires certain "skills according to the new rules") to obtain an explicit solution in situations where there is no traditional analytical solution for nonlinear equations or systems.



Will quantum computers take on conventional ones?

Published: #7 Autumn-2011

By Yu. V. Vladimirova, V. N. Zadkov, Physics Department and International Laser Center, M. V. Lomonosov Moscow State University, Russia



Nowadays nano-technologies allow mass-production of chips with a resolution of tens of nanometers, and even with atomic resolution in laboratories.

Elementary logic gates of these devices will soon be made up of tens or even single atoms, so that these devices will no longer obey the laws of classical physics, and their quantum nature will come into the play. Physicist Richard Feynman, a Nobel laureate, was the first who noted this fact in his famous speech at the dinner of the annual congress of the American Physical Society in 1959.

A quantum computer can be built of such kind of quantum logic devices. The question is what are its advantages and will quantum computers turn on conventional ones?

Classically in terms of information systems there can only be two possible logical states – 0 and 1, which are represented graphically in physics as states of the unit vector (up and down, respectively) on the unit sphere of states of such vector, which is called the Bloch sphere. As opposed to a “traditional bit”, a quantum bit of information (a qubit) can have not just two, but an endless number of states – since it is described by combination of two basis states, normalized to unity, $|0\rangle$ and $|1\rangle$ (analogues of states 0 and 1 of classical bit). The qubit state vector can end at any point on the Bloch sphere, including coinciding with value $|0\rangle$ (up) or $|1\rangle$ (down). The qubit “lives” in a state space, which is called the Hilbert space and is two-dimensional for a single qubit (two to the power of N for N qubits). There is remarkable and fundamentally important fact in terms of storing and processing information, that the quantum memory register (set of qubits) stores information more efficiently than any classical memory. Thus, N qubits “live” in a $2N$ -dimensional Hilbert space (for example, 2^{16} , i.e., 65536, for 16 qubits) and operations on them are carried out by turning vector of their state in this space in one step. As to conventional computer, $2N-1$ complex numbers are required for storing information about N qubits, and operations on them require resources, exponentially growing along with the number of qubits. In other words, when it comes to the advantages of quantum computers over conventional ones, the most important and actually the only one) is the multi-dimensional Hilbert space of quantum system states, which grows as $2N$ (N – number of qubits). In this sense, quantum computers can be classified as parallel computing machines.

Quantum computing

To describe the changes that occur with the state of quantum computing

systems, the so-called quantum computing language is used. By analogy with the conventional computer, which consists of electronic circuits containing wires and logical elements (LE), the quantum computer is built of quantum circuits consisting of communication channels and elementary quantum LE, allowing transmitting quantum information and manipulating it. By analogy with the classical LE, the simplest quantum LE are single-qubit element NOT and two-qubit elements. There is a so-called *completeness theorem*, which says that any multi-qubit LE may be composed of CNOT and single-qubit elements. The best known three-qubit LE is, for example, the reversible Toffoli LE.

An important question is whether any classical LE maybe compared with the quantum LE, as, for example, for NOR and CNOT? The answer is negative, because many of the classical LE are irreversible. (In 1961 Rolf Landauer, an IBM physicist, showed that even “perfect” conventional computers based on irreversible logic dissipates heat when it loses any bit of information. However, it is known, that the organization of computing with the help of the so-called conservative or “preserving” logic without destructing the information does not fall under the Landauer’s principle and allows creating power-efficient processors.)

For example, it is impossible to determine the input states by the output value of XOR, i.e. the information is permanently lost. To the contrary, the unitary quantum elements are always reversible, and, therefore, the result obtained at the output of the quantum LE can always be inverted by another quantum LE. That is why quantum circuits cannot always be used for direct modeling of the classical circuits. However, any classical circuit can be replaced by an equivalent one, which contains only reversible LE, for example, Toffoli LE. That is, the quantum Toffoli LE, as well as classical one, can be used to model

the classical irreversible LE. This means that quantum computers can perform any computation that can be realized on a classical computer. However this is not the advantage of quantum computers to conventional. The advantage is a high degree of parallelism of operations in the multidimensional Hilbert space of quantum computer states. The only question is whether such algorithms for quantum computing, which allows to realize this advantage, exist.

Quantum algorithms

Such algorithms exist, but they are few. They are all based on so-called *quantum parallelism*. There are three

classes of quantum algorithms having the advantage over known classical algorithms (see insertion above). The first one is based on the quantum Fourier transform. It includes the Deutsch-Jozsa algorithm, that is to define whether the binary variable function $f(x)$ is constant (0 or 1 on all inputs) or balanced (returns 1 for half of the input domain and 0 for the other half); as well as the Shor's algorithms for the problems of factoring and computing of discrete logarithm. Factorization time for numbers, represented by a different number of bits in the classical and quantum computers, is compared in the figures. The second class of algorithms is a quantum search

algorithm (the Grover's algorithm), which is the fast quantum algorithm for solving the search tasks in the space of N elements. The Grover's quantum algorithm can solve the search task in a time proportional to the square root of the time of executing the fastest classic algorithm, which is a great speed-up. Finally, the third class of algorithms covers quantum simulations, in which the quantum computer is used to simulate the quantum system itself.

How does a quantum computer operate?

The functional diagram of the operation of any quantum computer looks as follows. Its main component – the quantum register – is a set of a certain number of qubits. Before inputting information into the quantum computer, it is necessary to prepare the initial state of the register qubits, i.e. to perform *initialization*. As a result of this operation, all the register qubits must be transferred into the main basis states so that the state of the register can be written as $|01\rangle, |02\rangle, |03\rangle, \dots, |0N\rangle \equiv |01, 02, 03, \dots, 0N\rangle$. The initialization itself is a complex task. When the role of qubits, for example, is played by atoms or ions, then deep cooling (up to the order of micro-Kelvin) is required to transfer them to the ground state; and in the case of photons polarization methods shall be used. After the data register was set to the ground state, each qubit of the register can be put to an "excited state" $|1\rangle$ by selective acting to it (for instance, with the help of the pulses of an external electromagnetic field), and the entire register will therefore be transferred into a superposition of basis states, which determines a number in the binary system. The next step is to perform data input, i.e. to convert the state of the input register into a coherent superposition of basis states. This can be done, for example, by the means of the pulsed excitation of the system. In this form the information comes to the input

Size of modulus (bits)	1,024	2,048	4,096
Factoring time in 1997	10^7 years	3×10^{17} years	2×10^{31} years
Factoring time in 2006	10^5 years	5×10^{15} years	3×10^{29} years
Factoring time in 2015	2,500 years	7×10^{13} years	4×10^{27} years
Factoring time in 2024	38 years	10^{12} years	7×10^{25} years
Factoring time in 2033	7 months	2×10^{10} years	10^{24} years
Factoring time in 2042	3 days	3×10^8 years	2×10^{22} years

Conventional computer factorization time

Size of modulus (bits)	512	1,024	2,048	4,096
Quantum memory (qubits)	2,564	5,124	10,244	20,484
Number of quantum gates	3×10^9	3×10^{10}	2×10^{11}	2×10^{12}
Quantum factoring time	33 seconds	4.5 minutes	36 minutes	4.8 hours

Quantum computer factorization time

of a basic element of the quantum computer – a quantum processor that performs a sequence of quantum logic operations. The last step is to read the results, i.e. to measure the states of the qubits at the output.

Prototypes of quantum computers Building a working prototype of the quantum computer is one of the 21st century challenges for physics. There are already built prototypes that operate with tens of qubits, but the question of scaling of these devices still is open. The fundamental problem created by the Nature itself on a way of scaling quantum computing devices – is the problem of rapid decoherence of the quantum states of qubits due to their interaction with the environment. If one can eliminate this problem, the quantum computer could be scaled up to the size of the Universe, as the Universe is the largest of the known physical systems, and it obeys the laws of quantum physics. Recently S. Lloyd, a USA scientist, has calculated what the Universe is from the informational point of view. He estimated that the Universe can be considered as a quantum computer with a memory of 10^{90} bits, which had performed 10^{120} elementary logical operations during its existence (since the Big Bang). Many laboratories in USA, Europe, Russia, Japan and Australia have been developing various prototypes of quantum computing devices, based on the following key technologies: (i) solid-state quantum dots in semiconductors, (ii) superconducting elements (Josephson effects, SQUIDs, etc.), (iii) ions in the vacuum Paul traps (or atoms in optical traps), (iv) hybrid technologies (eg, photons in linear optical systems, photons in micro resonators, etc.). At the turn of millenium, single-qubit quantum processors were designed in many research laboratories. Soon afterwards IBM demonstrated a liquid NMR quantum computer (up to 7 qubits). In 2005 NEC, Japan, built two-qubit quantum processor based on superconducting elements. Around

the same time, prototypes up to ten qubits were demonstrated, based on ions and atoms in traps (USA, Austria, Germany). In 2009, researchers from Yale University (USA) created the first simple solid-state quantum computer – two-qubit superconducting chip, which was able to perform simple quantum algorithms. A team of scientists from Bristol University (UK) also created a semiconducting chip for quantum computing based on the principles of quantum optics and involving conventional optical elements (mirrors, refracting plates, etc.). The Shor's algorithm has been demonstrated with the help of this chip. In February 2007 D-Wave Systems in Canada, (www.dwavesys.com) introduced the Orion - the first working prototype of quantum computer, based on 16-qubit chip. In December 2008 the company officially launched the project of distributed computing based on the adiabatic quantum algorithms implemented in adiabatic superconducting quantum computers D-Wave. In May 2011 the D-Wave One was presented. It was based on superconducting 128-qubit processor chip, housed inside a cryogenics system within a 10 square meter shielded anti-magnetic room (to eliminate the influence of external magnetic fields). The company plans to scale this chip to 1024 qubits, and signed a contract with Lockheed Martin Corporation for the development and application of such quantum processors to meet the

needs of the corporation. In Russia, several research groups of the Physical and Technological Institute of the Russian Academy of Sciences, Lebedev Physics Institute of the Russian Academy of Sciences, Institute of Spectroscopy of the Russian Academy of Sciences, Institute of Laser Physics of the Russian Academy of Sciences, M.V. Lomonosov Moscow State University, Moscow Physical-Technical Institute and various other institutes are involved in the physics of quantum information and development of prototypes of quantum computers and quantum computing devices. However there is still a long way to go concerning the creation of a valid quantum computer. In the next 20-30 years working prototypes of quantum computers with practically significant number of qubits, as well as specialized quantum computers similar to those developed by D-Wave Systems, are likely to be created to perform quantum algorithms for solving real problems, first of all, of quantum physics, materials science, etc. One should note that when such quantum computers are created, they will become more effective than conventional computers only in solving a limited class of problems that we discussed above. So the answer to the question in the title of the article - will quantum computers turn on conventional one? – is negative. No, in the foreseeable future they won't. ■■■

Quantum computers can perform any computation that can be realized on a classical computer. However this is not the advantage of quantum computers to conventional. The advantage is a high degree of parallelism of operations in the multidimensional Hilbert space of quantum computer states

From «Dragonfly» to the «ZETA SCALE»

Editorial
interview

Published: #9 Spring-2012

Our editor Igor Levshin talks to one of the most respected experts in the field of HPC, Steve Scott, at NVIDIA's GPU Technology Conference in Beijing (GTC Asia).



Steve is the chief technology officer of the Tesla business unit at NVIDIA. He joined NVIDIA after 19 years at supercomputing company Cray Inc., including the last six as CTO.

Igor Levshin: The effectiveness of a supercomputer is determined by connections. At Cray networks you developed Giga Ring, the Red Storm, and most recently the Aries network in the Cascade project. Which topology and concepts are up to date and which of them will be most successful in the future?

Steve Scott: The Jaguar system at Oak Ridge National Labs in Tennessee used the Seastar 2+ network, with a 3D-torus topology. The follow-on network, Gemini, also implements a 3D torus, and will be used in the updated Titan system later in 2012. The next generation Cray network, Aries, implements a "dragonfly" topology, and will probably appear in late 2012 or early 2013.

I. L.: Will more sophisticated topologies be required in the future? The creators of K Computer describe its topology as 6D-torus. Will the dimensions increase in the next generations of networks?

S. S.: I think the future trend for interconnects will involve reducing the diameter of the network. At the same

time, the router will provide higher bandwidth than it does today. Dragonfly allows you to reduce the network diameter, scaling to large system sizes with the maximum distance between processors of five hops. Increasing the number of dimensions in a torus will also reduce network diameter, though not as significantly. In the immediate future, shorter network connections will be made from copper, and longer ones will use optical technology.

I. L.: An average Supercomputer Centre is differentiated by a much higher density of heat dissipation. Everything must be geared towards reducing slowdowns. Will this key issue be removed if the connections are optical?

S. S.: Optical links will relax the need to pack components closely together, which can help heat dissipation. But we can't just use optics everywhere. Unlike cloud servers which typically use optical interconnects, supercomputers require very high bandwidth interconnects, which means more cables. And this is a question of money: as a rule, the shorter the distance and cable requirement, the cheaper the solution. There are other ways to save money in the network. For example, as we approach the exascale era, the network interfaces will be integrated

onto the processor chip, which will make the network more performance and energy efficient.

I. L.: Can you tell more about integration? What are its limits?

S. S.: Integration in future processors has three aspects.

The first aspect is the integration of the memory. A 3D-stacked memory is the first step in this direction. In this case, memory is located on the package with the processor, if it's not integrated onto the processor chip. Bandwidth will increase and energy consumption will be reduced.

The second aspect is integrating the GPU and CPU onto a single processor. Up to 90 percent of the chip can be used for GPU (throughput optimized) cores for parallel processing, which will be fast and energy efficient. But it will still be necessary to have several traditional serial CPU cores optimized for fast single-thread performance. It makes sense at some point to integrate them into the silicon, or at least into the same package. Finally, the network interfaces of interconnects will be integrated into a single chip. These heterogeneous systems, similar to those which are now implemented in nodes and networks, will be implemented on each processor

chip in future systems. There will be thousands of cores in the processor, they will be connected by an internal high-speed network, and their external network will be connected through fast interfaces. It will happen as we approach the exaflops era. These cores may be equipped with an ARM instruction set, as ARM is rapidly breaking into the world of HPC. Explicit execution of operations such as load/store in global address space is required, and the path of the package will be automatically determined by the interface: whether it is communication within the chip or the need to refer to another node.

I. L.: What should I think about when writing an application using the capabilities of the GPU?

S. S.: The most important thing is to ensure the application is enabled to take advantage of parallelism. If it the algorithm exposes the parallelism, the compiler and runtime can then map it to a specific architecture. When we launched CUDA, it was very important to give programmers a tool for high-level programming. Today, they can use directives in C and FORTRAN programs to more easily expose parallelism in the application. The programmer is now able to «suggest» to the compiler where to extract the parallelism and what can be executed on the GPU cores. But the same code will still work on the CPU (although not as fast). Now the programmer can spend time on exposing parallelism in an application, rather than creating software versions optimized for specific architectures. The hardest thing for a compiler to do is detect where you can parallelize application code. This can require whole program analysis, and can be made difficult due to pointer disambiguation. But when the programmer can tell the compiler «this is safe to run in parallel, everything will be fine», while the compiler and runtime take care of the low-level implementation of parallelism, it's much easier. We've found that directives work well. Typically, performance using directives and the compiler can approach that obtained manually by a programmer. While using directives for most of the

code, you can still get control at a lower level using CUDA if you want.

I. L.: Another smart part of the compiler is to understand the avenues of internal or external networks. What is the situation here?

S. S.: Of course, this is also very important. Here, NVIDIA has developed and optimized a technology called GPUDirect to help speed communications. GPUDirect allows direct access to the GPU memory, between GPUs in a single node or across the system interconnect, via InfiniBand for example, while bypassing the main system memory. Inter-node communication is particularly important for the scalability of large systems. Communication latency and overheads are especially critical when we're talking about the global address space languages. To optimize for load/store access across the system, you need a more closely integrated network.

I. L.: How will the systems of the exascale era «understand» which compute nodes we should store data close to, and where to look for the data a compute node needs? Who would take over this intellectual effort: system software or application?

S. S.: The distribution of data is probably the biggest challenge for parallel computing. Most of this work will be handled by the applications themselves. Most scalable applications will use explicit parallelism. Using MPI or PGAS, they will decide which data to be processed locally and how to minimize communication with remote nodes. This sort of data distribution is inherent in the programming model.

There still remains the problem of choosing in which level of local memory hierarchy the application should store data. This problem is becoming more and more important because data movement is becoming increasingly expensive compared to computation: adding or multiplying two numbers takes far less energy than moving the operands across the chip. The locality of data within a node is no less important than locality at the network level, and this task should be explicitly managed by software.

With 3D-stacked memory, we've added another layer to the traditional main memory. The compiler will have to decide what to keep in the stack and what should be stored in the main memory. The compiler probably will often be able to do this job, but perhaps the programmer will need to expose information regarding the locality of data – similar to directives for parallelization. And the compiler will handle mapping it to a given system configuration.

I. L.: Are the efforts of energy efficiency enough to meet the «1 exaflops per 20 MW» principle?

S. S.: Currently, even for the most efficient processors this ratio is dozens of times less than needed, but not thousands as it was some years ago. The most efficient processors available today are around 1 GF/W. It's expected that an exaflops machine will consume 20 MW, that means we need to improve energy efficiency 50 times. Part of the gain will be achieved by improvements in silicon technology. By 2018, we will have IC technology with a feature size of about 10 nanometers, but the threshold voltage will only be dropped a bit, limiting the power efficiency improvements and leaving us far short of our goal. The rest will have to come from more sophisticated architecture and circuit design. I'm not sure we'll be able to meet 20 MW, but I do think it's realistic that we will meet 30 MW.

I. L.: Is NVIDIA making estimations about the energy efficiency of exascale era processors already?

S. S.: Yes we are. We're been performing analysis as part of our Echelon project under the DARPA UHPC government program.

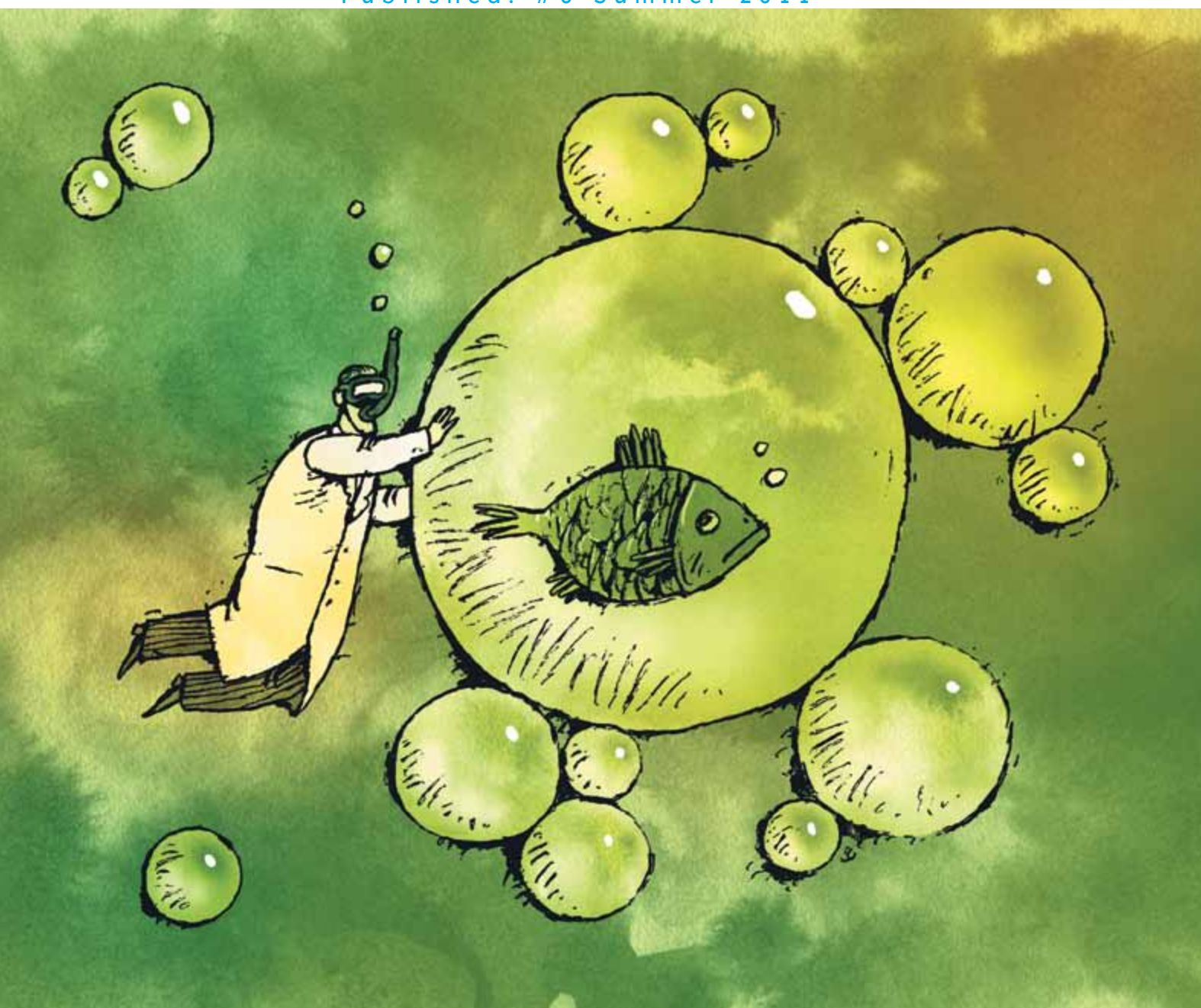
And we've been looking specifically at developing processors for exaflops machines.

And NVIDIA is not the only one thinking about exascale. Now everyone's talking about exascale: the US, China, Europe, India, and Russia are as well. It's very clear that zetascale computers will not be built using silicon transistor technology. What will the future bring? Who knows! Let's get back to this conversation in 2020, OK? ■■■

Eddy bearings and liquid balls: virtual and real

By B. Zotov, I. Drozdov, A. Wagner, A. Pozdnyakov, N. Vladimirova
Illustrated by Vladimir Kamayev

Published: #6 Summer-2011



Human civilization followed by pumps development as one of the main technical equipment. What kind of innovative solutions could be found in the science fields like pumps development, which has a thousand-year history.

Hydro(aero)dynamics is one of the most difficult and interesting fields of contemporary physics. Considerable knowledge, experience and intensive development of hydrodynamics was accumulated during last four centuries. A centrifugal pump was invented by Denis Papin in 1649! Jacob Bernoulli published his first work "Hydrodynamics" in Strasbourg in 1738. It made occurred for people that all problems in this science must have already been solved for the last centuries.

15 years ago a new field in hydrodynamics – eddy hydrodynamics started its intensive improvement. It facilitates the efficient use of eddy structures in technical equipment. The difficulty of the physical processes is described by the system of equations, numerical solution of which requires more RAM than the memory of a modern computer. The necessity of finding of optimal correlations of governing parameters for a technological solution may be very costly because

it requires manufacturing and testing a significant number of test specimens. For this purpose usage of a supercomputer with its enormous memory may considerably reduce a number of experimental tests and lead to discovery non-evident effects. Here are two examples. Bearings can be of two types – ball and sleeve. Sleeve bearings have a thin liquid layer. Can sleeve bearings be equipped with 'liquid balls' eddy structures? Yes, they can! This is especially important for sleeve bearings with low-viscosity fluids in order to enlarge the bearing clearance. For this purpose a structure with cavities of a particular profile should be created on the surface of a bearing liner and/or bushing. The effect of additional forces has already been proven by the numerical solution of a single cavity flow problem. Solving the problem with the help of a supercomputer may optimize the number and geometric parameters of the cavities. Another example is a screw propeller. It looks like that everything

has been already studied and that it is nothing is left to discover. However, this is not true, especially for heavy-tonnage vessels with low-speed large-size propellers requiring the use of large-scale low-speed engines. In the case of high-speed engines, for example, steam turbines installation of bulky reducers is necessary.

The limitations on the propeller speed result from the following main reasons:

- outflow from the blade edges that increases with the rotations speed decreasing the propeller efficiency;
- insufficient rigidity of the propeller blades.

Can a technical solution 'kill two birds with one stone'? Yes, it can! It is just necessary to make a special profile of the blade trailing edge.

State-of-the-art applied and fundamental flow dynamics and deformable media problems are solved by means of high-level discretization of the physical space with meshes varying from several million to several billion cells, depending on the nature of a problem. Only supercomputers and high-efficiency (HPC) cluster systems are capable for the solution. Normally, HPC systems are built on blade platforms consisting of dozens and hundreds of nodes (multi-nucleus processors) resulting in thousands of nuclei. The total capacity of such clusters now varies from dozens of teraflops to several petaflops; they have hundreds gigabytes of random access memory. A reasonable computation time varying from a few minutes to a few days and required high accuracy of the results in fluid dynamics can be obtained only with these powerful computers.

State-of-the-art hardware-software complexes consist of supercomputer clusters and cutting-edge computational technologies and Computational Fluid Dynamics (CFD) numerical techniques. They are used for the solution of

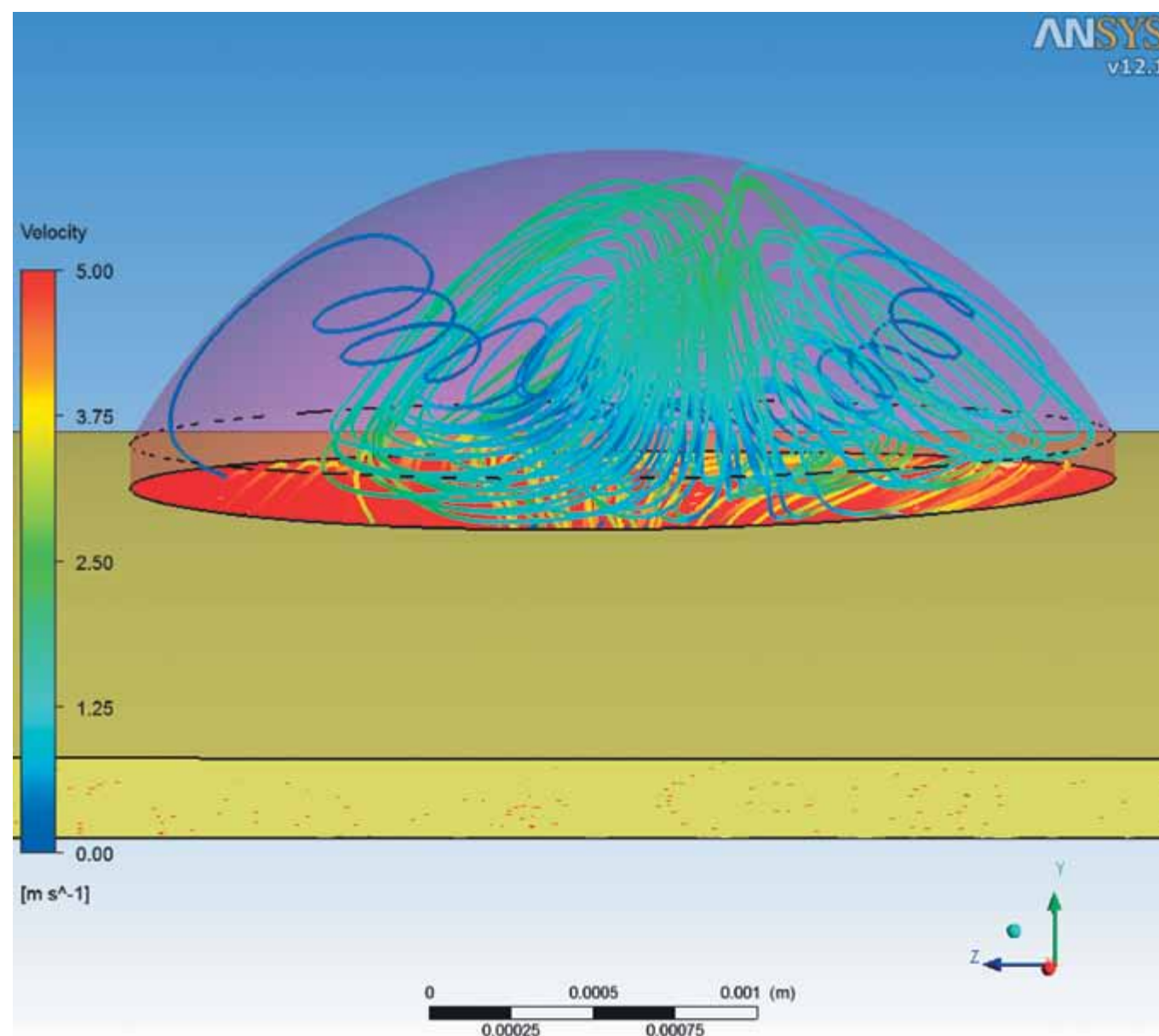


Fig. 1. «Liquid eddy ball» – local eddy structure in the bearing hemispheric cavity

various scientific and industrial hydrodynamics problems (one-phase and multi-phase internal fluid and gas flows in channels, pipelines, flow rate meters, pumps, external flows of ships and vessels, sports yachts, submarines and deep-submergence vehicles, screw-propellers, etc.) It is important that these problems are quite well parallelized into several

dozens or hundreds processes. Computer-based modeling is usually performed on the basis of in house codes or commercial CFD packages (for example, ANSYS CFX/FLUENT, STAR-CD, STAR-CCM+). Codes perform numerical integration of Reynolds-averaged Navier-Stokes fluid and gas dynamics equations (RANS/URANS).

The problem of development of sleeve-type bearings for low-viscosity fluids by creating periodic eddy structures, or so-called “liquid balls” is one of the most challenging tasks. The idea is to create a regular eddy structure of a particular profile on the surface of a bushing in a thin lubricant film between the liner and the bearing shafts. The structure

Considerable knowledge, experience and intensive development of hydrodynamics was accumulated during last four centuries. A centrifugal pump was invented by Denis Papin in 1649! Jacob Bernoulli published his first work “Hydrodynamics” in Strasbourg in 1738. It make occurred for people that all problems in this science must have already been solved for the lasts centuries

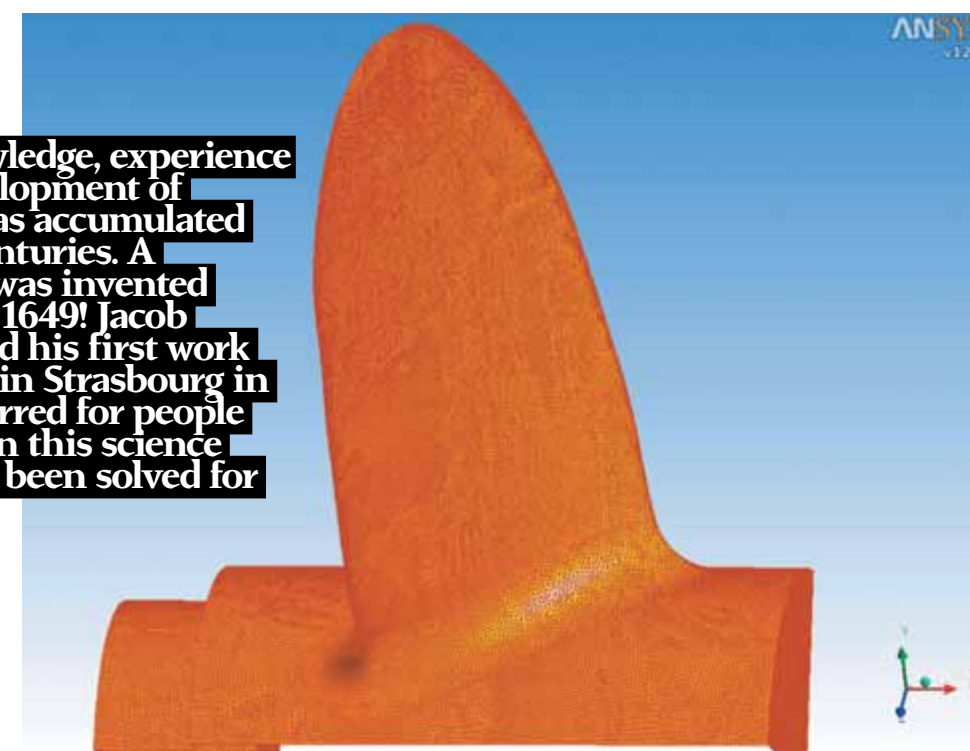


Fig. 2. Fragment of a surface mesh on a blade

creates additional hydrodynamic forces. Supercomputer numerical modeling proved the expediency of this idea. Fig. 1 shows the calculation results – spatial streamlines proving the availability of local eddy formations in the area of a semispherical cavity located on the boundary of a very thin (0.1-0.2 mm) lubricant layer around the rotating bearing shaft. Such multi-scale hydrodynamic effects can be simulated only at clusters with possibility of efficient task parallelization into dozens and hundreds of subtasks. To understand to the supercomputer’s role in such calculations better, let us provide some numbers: the lubricant layer thickness is 0.2 mm, the diameter of the rotating bearing shaft is 100 mm, the shaft rotation speed is 3,000 revolutions per minute, the diameter of a ‘liquid eddy ball’ is 2 mm;

the distance between ‘eddy balls’ (cavities) is also 2 mm. To provide an adequate reproduction of such periodic liquid flow structures in the ultra-thin lubricant film of a bearing – sufficient discretization of the domain is necessary. Mesh sizes of about several dozens and hundreds million cubic cells are required. The calculation results shown in Fig. 1 were obtained on the grid of 3.5 million of cells falling within a single cavity. These results have proven previously made assumptions about the effectiveness of such design. They will serve as a basis for the creation of new equipment with sleeve-type bearings for low-viscous fluids. Supercomputer simulations of screw propellers with profiled trailing edge blades were completely proved by water-tank experiments. This example clearly shows that numerical

modeling can practically completely replace full-scale natural experiments. In this case, the calculations provide much more diverse information about the structure of the flows and distribution of hydrodynamic forces, eddy formations in flows and other local features of the flow. A marine screw propeller with a standard blade and a propeller with a specially profiled trailing edges were investigated. It was necessary to calculate the velocity and pressure fields for the blades of such propeller; calculate the propeller momentum and compare the specifications of such propeller with the propeller parameters with straight non-profiled trailing edges. First calculations have shown that the specially profiled trailing edges of a propeller blade lead to the destruction of the eddy structures behind the propeller.

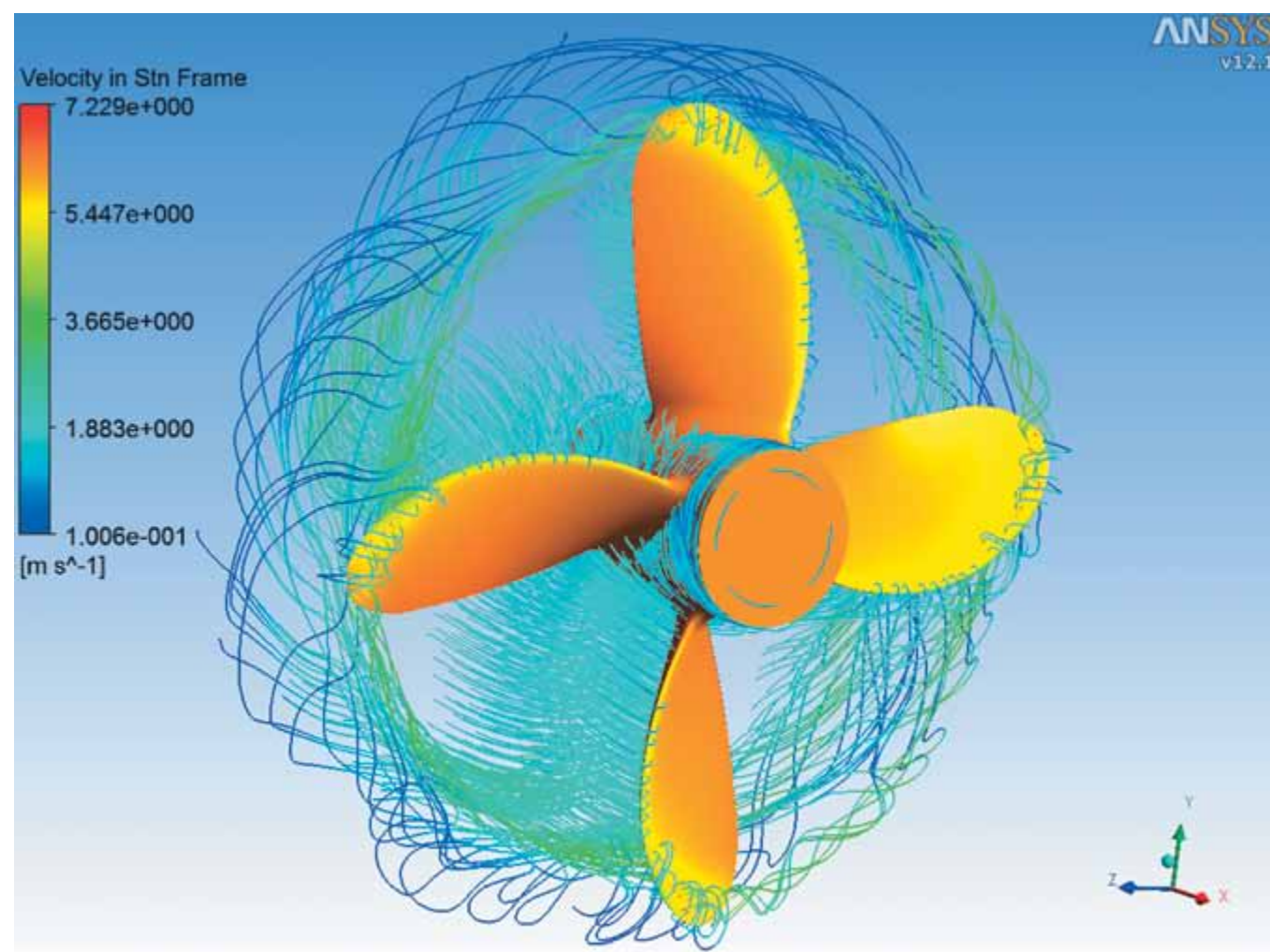


Fig. 3. Streamlines in the vicinity of a propeller

Hydrodynamic and acoustic pulsations are getting lower; the momentum of the propeller increases. All calculations were performed on the cluster ensuring the discretization of the hydrodynamic flow domain behind the propeller with a mesh containing around 50 million cells. Fig. 2-3 show the fragments of the surface mesh on the blade and the streamlines in the area of propeller blades of standard design. Things are quite different for the blades with the profiled trailing edge. Some specialists were doubted the effectiveness of such design. However, performed comparative water tank tests of the designed propellers with standard trailing edges and specially profiled propellers have proven the efficiency of such propeller. ■■■



Hadrons and human health

By Sergey Nemnyugin, Sergey Merz, Olga Ruban

Published: #8 Winter-2011-2012

Computer modeling could make the treatment of cancer more effective. The role of the technology regarding supercomputer technology in this treatment can't be overestimated.

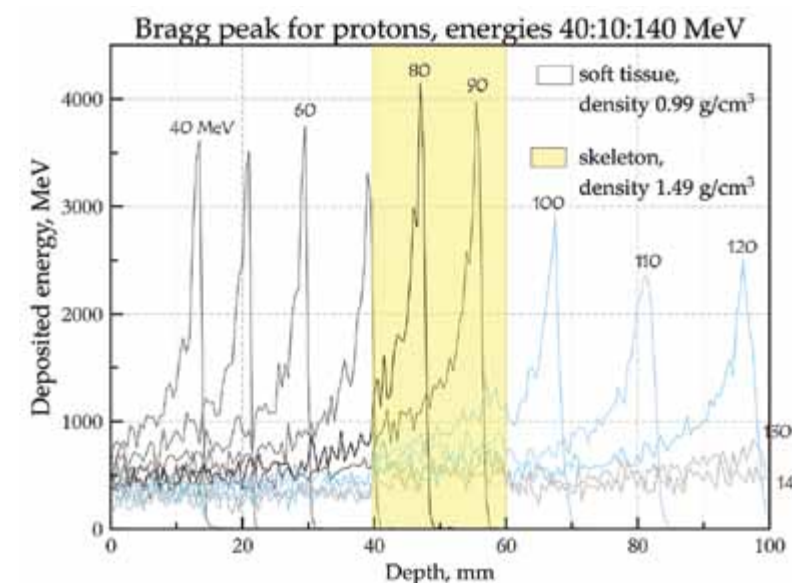
Treating cancer includes three approaches – surgery, chemotherapy or radiation therapy. Radiation therapy is based on using ionizing radiation to damage DNA strands in cancer cells. As a result, cancer cells lose their reproductive ability. In some cases, it results in the complete cure of human organisms. Traditionally, radiation therapy uses gamma radiation produced by cobalt isotopes' decay. Gamma radiation penetrates the tissues of patient, simultaneously effecting healthy cells and biological tissues. This is one of drawbacks of the traditional approach.

Exposure of the affected organ not by gamma-radiation but by beam of protons or carbon nuclei overcomes this disadvantage. This approach is called «hadronic therapy», it has several advantages and is more prospective in medicine. The main advantage of this method of treatment is the exact direction of the energy impact to the affected area, and minimal damage to healthy tissue. Based on this physical effect, called «Bragg peak» the dependence

of absorbed dose regarding the penetration depth has a narrow peak at certain depth, and position of this maximum is controllable (Fig. 1). Obviously, the nature of the absorbed dose profile of heavy charged

particles allows us to perform precise exposure of malignant tumors, thus severely reducing the ionization of surrounding healthy tissue, as compared to gamma ray exposure. This exposure method has been

Fig. 1. Profiles of absorbed doses (Bragg peaks) in a simple heterogenic biological model (composition: alternating soft and bone tissues)



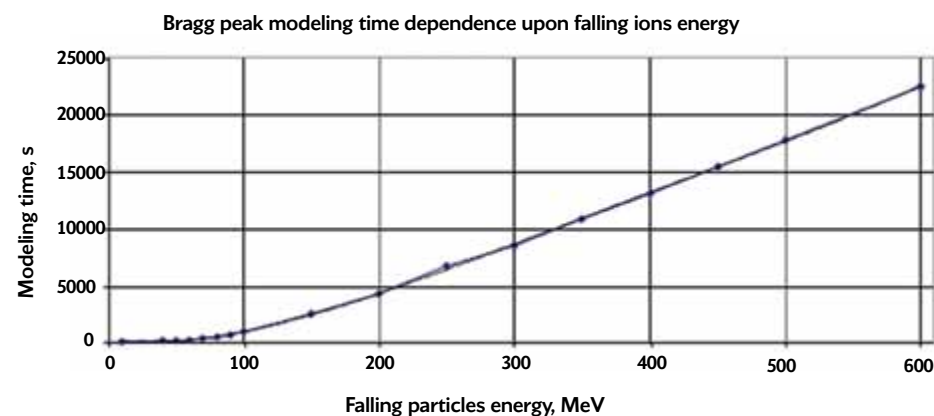


Fig. 2. A time-effect plot modeling a Bragg curve of carbon ions in water – depending upon transmitted bundle energy

actively developed since the mid-1950's, its effectiveness is confirmed by numerous clinical trials. At present, several dozen hadron therapy centers are in operation around the world and the number of successfully cured patients exceeds 30,000. There are many problems in this area whose solution requires using computing technology – the design of particle accelerators and equipment centers, the development of treatment planning methods, simulation of radiobiological effects of exposure, etc. All these problems require both computing resources and appropriate software. The most widely used software packages are Geant4 and FLUKA. These packages use Monte Carlo methods to simulate particles/substance interaction.

Optimal exposure strategy can be chosen by simulating the complex processes of beam/biological tissues interaction, taking into account the inhomogeneity of chemical composition and physical characteristics as well as the complicated dynamic (i.e. changing over time) geometry of inner organs. This kind of modeling can take from several minutes to several hours, depending on the energy of the incident particles and their number and complexity.

The energy of the particles plays an

important role, since this parameter is responsible for the penetration of particles into substance. At the same time the duration is determined by the energy calculations. E.g. Fig. 2 shows the «Bragg curve of simulation time dependence» on the energy of falling carbon ions. The curve shows

that even for the medical energy range (200-300 MeV), computation takes two hours, even by modeling a single curve. In reality, the exposed tumor is a volume, not just the point. This implies some difficulties. Fig. 3 shows schematically the simple geometry of exposed area and dosage profiles for different sorts of radiation particles. Using gamma-ray leads to the undesirable exposure of tissue as both the entrance channel and the critical organ are located just outside the tumor. Proton therapy minimizes the exposure of critical organs, but the input channel tissue continues to uptake a significant dosage. Applying ion therapy allows to clearly localize the maximum dosage area, minimizing the exposure of critical organs and tissue. However, due to the fragmentation effect of carbon nuclei there is a «tail» of absorbed dose just behind the tumor. This «tail» leads to exposure of critical organs. To achieve the «plateau» effect, one should scan the tumor by a narrow,

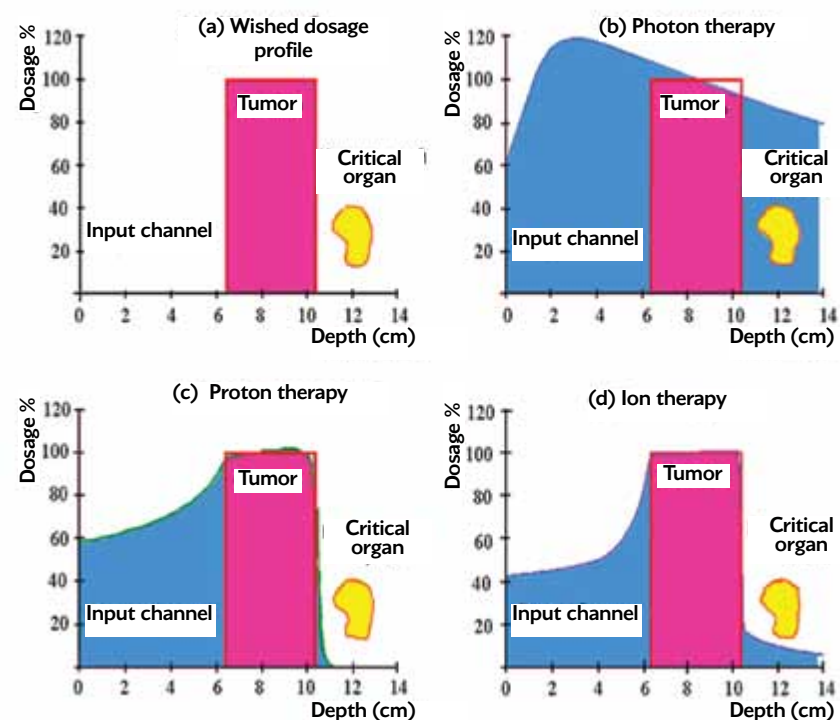


Fig. 3. A geometry scheme of exposed areas and profiles of various particles

directed beam of particles, changing its energy and direction. Fig. 4 shows the results by simulation package FLUKA. The resulting curve is called the modified Bragg curve. In order to scan the area about 4 cm, up to 50 measurements are required with ion energies ranging from 150 to 199 MeV. To optimize computing technology, OpenMP was used. The complicated geometry is yet another reason for prolonged calculations. The human body is a very complex structure in terms of geometry, physical and chemical composition. Creating a model accurately describing the complexity of human body is a complex task. Inclusion of such models into the calculation of the Bragg curve seriously increases the run-time of modeling.

Fig. 5 shows the sample calculations of the Bragg peak for a proton beam passing through a cylindrical structure consisting of biological tissues. The calculation is made by Geant4 package. The high-performance computing technology may be applied to the calculation of the localization depth of Bragg peak and its «fine structure» depending on the range of hadrons's beam parameters (in

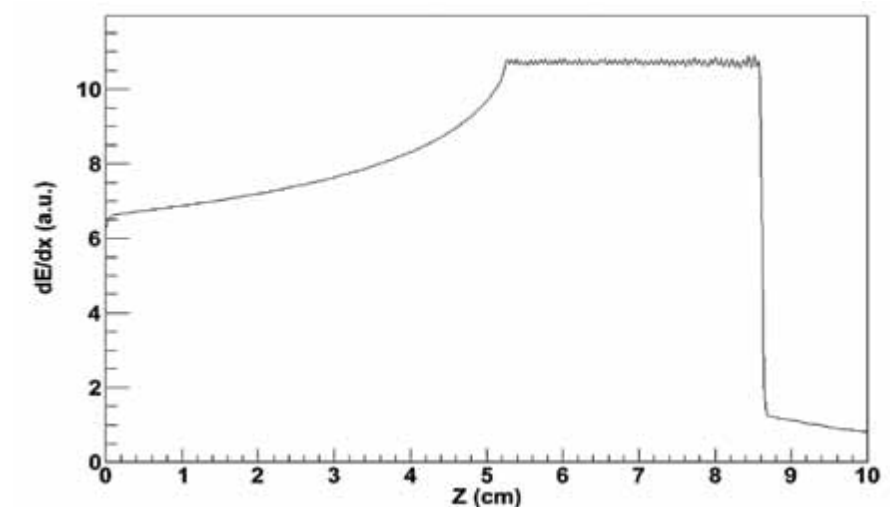


Fig. 4. A modified Bragg curve for carbon ions

which energy is the most important parameter). This allows to select the optimal set of parameters (energy). In this case there is a «trivial» data parallelism, when solving a large number of independent tasks. The target platform can be a distributed system. Its effective use is possible with a uniform and balanced load of computational nodes (but there is some difficulty to do this). As noted earlier, calculation time may vary depending on the hadrons

energy beam, differing by orders of magnitude, so to minimize the calculation time efficient algorithms for scheduling tasks are required on the system computing nodes. The second «dimension» of parallelism is due to complexity of calculations, and above all, the complexity of models used.

This case requires using parallel computing algorithms. For numerical solution of problems associated with the spread of elementary particles and electromagnetic radiation in different media, different variants of commonly used statistical modeling (imitation) are used. Statistical modeling algorithms get paralleled in a fairly obvious way, and parallel algorithms scale well. We should note, that implementing parallel Monte Carlo algorithms bring a variety of technical problems, i.e., the problem of the efficient generation of large number of independent streams of pseudorandom numbers with good probabilistic properties. Additional complexity is addressing the real and complex geometry of the exposed body, uneven physical and chemical properties of biological tissues, as well as a dynamic, flexible nature of their geometry.

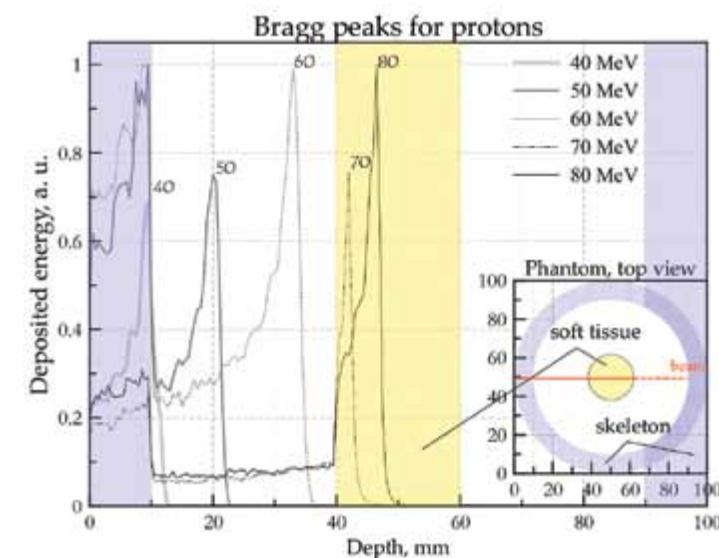


Fig. 5. An example of calculating Bragg peaks for complicated geometry

Modeling of biological reactor – project «artificial liver»

According to Colin McGucklin, professor of Newcastle University, engaged in solving the issue of creating the «artificial liver», the possibility of piecemeal «repair» of this vital organ will appear in the next 10 years. «Patches» will be made of stem cell-derived material. And in the future – in this regard professor hesitates to specify time even approximately – humans will be able to have the whole liver replaced. However, at the present moment the reality is that medicine involves «external devices» to help patients with damaged liver, as it has long been used in cases of acute renal failure.

But this task is much more complicated than hemodialysis – where a solution of electrolytes is used as the working fluid for cleansing blood. The liver hosts a more complex processes to cleanse the blood of toxins, poisons and allergens. The «Chemical» detoxification of blood in an «artificial liver» through a «cocktail» of chemicals is not possible, since it is also required to have a synthetic function – the creation of proteins and other complex compounds that are vital to the body.

At present, various research centers develop mechanisms to reproduce the cleansing function of the liver with the help of the biological reactor. The biological reactor is a series of stacked plates with live cell cultures, which are «stitched» by permeable capillaries that feed the blood plasma or serum that contains toxic substances. Purified plasma returns from the reactor back into the patient. Both «prepared» swine liver cells and artificially stem cell-cultivated cells can be used as a cellular structure. There are several reasons why it is impossible to use «artificial liver» endlessly. However, this blood cleansing technology allows saving many patients who are awaiting a donor liver. Also, it can seem that patients own livers can be significantly restored when connected to such an artificial device.

Works on creating «artificial liver» are currently being carried out at the Privolzhsky District Health Center (POMC), as part of the Federal Medical-Biological Agency – located in Nizhny Novgorod. There is no biological reactor, which would pass clinical trials, yet. However, the most important task has been already solved – computer modeling of the processes occurring inside the reactor was performed, due to which its optimal design was developed. The Reactor is a complex system with a number of processes occurring: biological, chemical, hydrodynamic. Moreover, these processes take place in a large volume of the working area

of the reactor. Therefore, the modeling required to employ great computer power, which is impossible to reach by using even the most powerful personal computers.

POMC acted in the capacity of task originator. Two subcontractors were attracted to work: small innovative enterprise «Scientific research center of special computing technologies», Nizhny Novgorod («NITs SVT» LLC) and well-known in the computer world's «T-Services», part of «T-Platforms» holding, which provided the project with supercomputer resources.

Microwave gun that fights malignant tumors

Along with this development, the same three companies have been solving another issue relating to the practical oncology. As part of this task, the influence of high power pulsed electromagnetic radiation on malignant neoplasms in parenchymatous organs (liver, spleen, kidney, pancreas and thyroid gland) and on the walls of hollow organs of the gastrointestinal tract was modeled, with developing methods of calculating the power area of assured destruction of the tumor in a variety of biological tissues, including influence on the surrounding healthy tissue to prevent the preservation of the viability of cancer cells.

Difficulties in solving this issue are pre-determined by complex dependence of the temperature reactions in a variety of biological tissues on changes in regional blood flow, high-energy impact and several other parameters.

Both these issues were solved on the basis of Cublic (R), software platform designed in NITs SVT, which provides the ability to create computer applications of almost unlimited complexity, and may be used for solving set tasks by a specialist, who knows the mathematical apparatus of his/her subject field, but who has neither programming experience,

nor even programming skills. This is achieved through the implementation in software of modern approaches and techniques both in terms of user-friendly interface and data and computation management. Mathematical models describing both the impact of microwave pulses to living matter and the behavior of the cell mass in the bioreactor, as well as multi-directional hydrodynamic processes in the bioreactor of «artificial liver», are based on the method of cellular automata. This discrete method, being effectively applied in mathematics, mechanics, physics, biology, chemistry, considers the matter as a regular lattice, where

each node has two stable conditions. With regard to living matter, these two conditions mean that the cell is either alive or dead. There are also two binary oppositions, describing the dynamics of the process: the cell diffuses / does not diffuse and is divided / not divided.

It would be possible to use classical equations of mathematical physics for set tasks, but for such a complex structure, which biological tissue has, it is poorly applied. This is because the original task of the numerical solution of one differential equation with certain boundary conditions breaks down into a large number of similar equations with the alignment of the

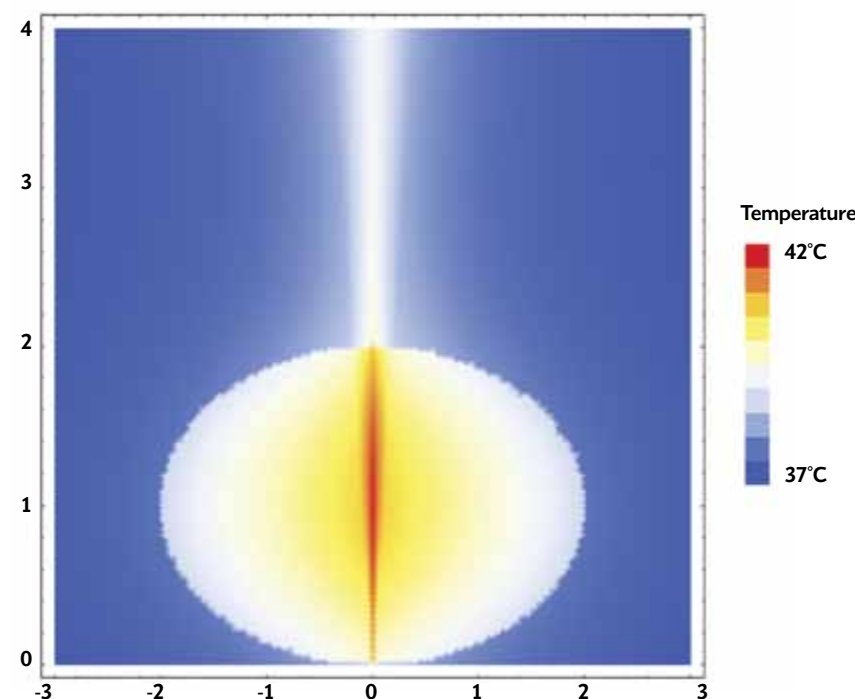
In the future humans will be able to have the whole liver replaced.

Supercomputer in theoretical medicine

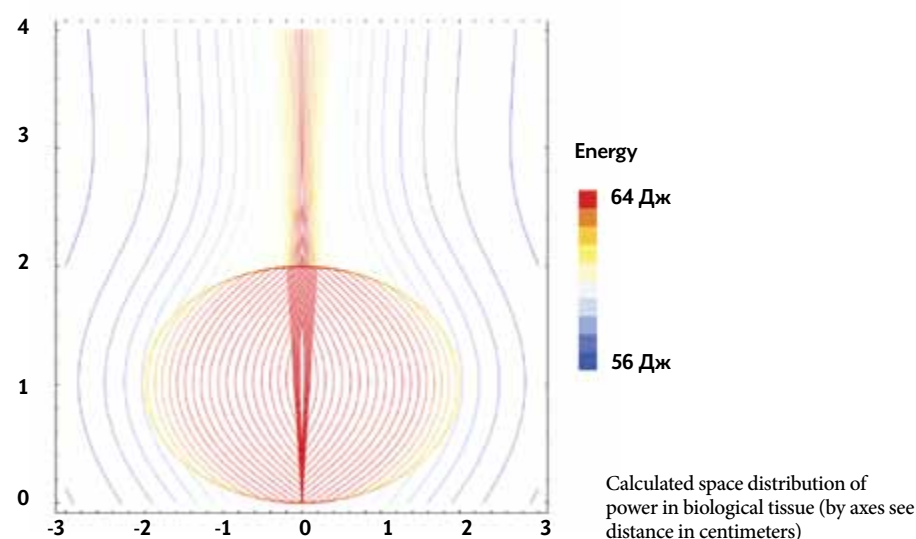
By T-Services, Vladimir Tuchkov

Published: #6 Summer-2011

Humanity is rapidly moving on the way of creation of a full set of spare parts that are able to replace defective human organs. Privolzhsky District Health Center is also working to address this global problem. It is a large multi-field institution engaged in both perspective scientific research and practical solving of issues of oncology diagnostics and treatment, transplantation.



The field of temperature distribution in biological tissue (axes show distance in centimeters)



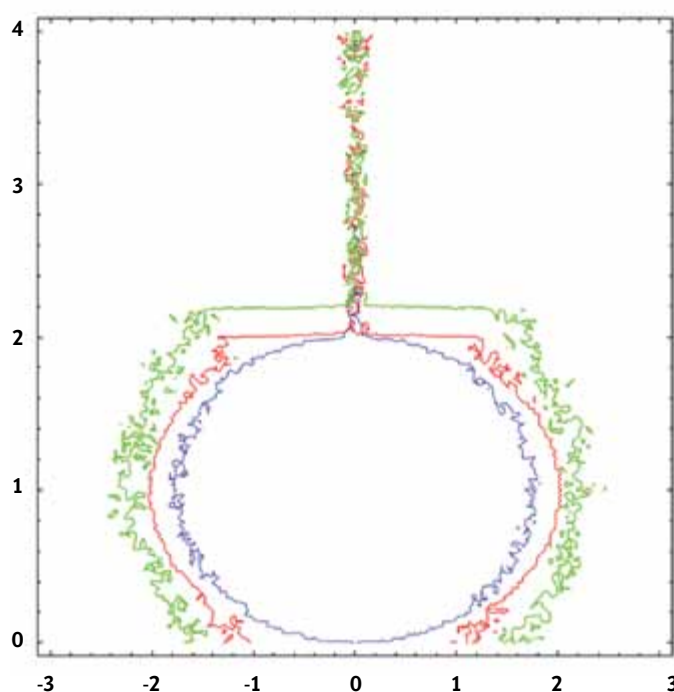
Works on creating «artificial livers» are currently being carried out at the Privolzhsky District Health Center (POMC), as part of the Federal Medical-Biological Agency – located in Nizhny Novgorod

PROJECT PARTICIPANTS

POMC – Federal Budgetary Health Institution «Privolzhsky District Health Center» of the Federal Medical-Biological Agency of Russia. It is a large multi-field institution, consisting of a network of hospitals, located in the Privolzhsky and Central Federal Districts. In addition to the standard set of diagnostic and treatment services, the Center provides high-tech medical care, which primarily includes cardiac surgery, liver and kidney transplantation, neurosurgery in case of brain tumors, treatment of inherited and systemic diseases, severe endocrine pathology, surgery with a high degree of complexity, including those in case of cancer. The Centre has a sufficient technological base and unique personnel resources for the successful scientific research, further developing of new medical techniques and setting tasks for the design of effective medical equipment.

boundary conditions for each layer of the tissue. This greatly complicates the program implementation and multiplies the complexity of the algorithm without increasing its accuracy. Modeling with cellular automata allows to eliminate this effect and to achieve significant paralleling of computing algorithms. In modeling of biological processes by means of cellular automata both in case of the microwave exposure, and in the bioreactor, Gillespie's stochastic algorithm was used, which have been widely used in recent years, especially in chemistry and biology. This algorithm allows for more efficient use of a probabilistic model of division, multiplication and diffusion of cells. In the modeling of hydrodynamic processes within the bioreactor, NITs SVT used the Lattice-Boltzmann method, which is a discrete model of continuum and is usually used in high-performance computing. The requirements of modeling with the use of supercomputing is to discover the processes in real time.

Calculation of the distribution zones of tissue necrosis. Blue – the border of zone of total destruction, red – the border of the border zone, green – channel for the input of working fluid of microwave emitter into biological tissue.



«**T-Services**» – a company providing a full range of services in the market of high-performance computing: from the leasing of Supercomputer Center's computing powers to the complete cycle, starting from the formulation of the task and ending in receiving final results.

The company implements computer projects in a wide range of industries and scientific fields, including petroleum and shipbuilding industries, computer graphics, biotechnology, engineering, nuclear physics and biochemistry. «T-Service» has its own Supercomputer Center, which is built on the basis of different platforms and configurations, connected by a high-speed 100Mb/s network with M9 point of international traffic exchange. The nodes with the following configurations are provided to the customer:

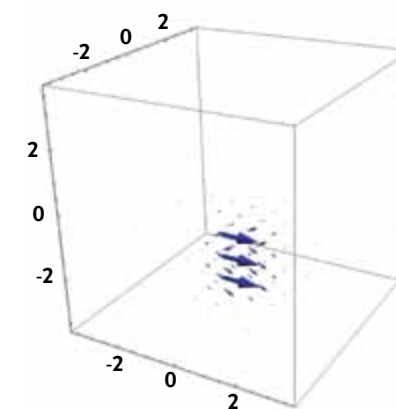
- 2 x AMD Opteron 6174 «Magny-Cours» 2,2GHz 64 GB RAM DDR3.
- 2 x Intel Xeon E5450 3.0 Ghz 16 Gb RAM DDR2.
- 4 x AMD Opteron 6174 «Magny-Cours» 2,2GHz 256 GB RAM DDR3.
- 2 x Intel Xeon E5450 3.0 Ghz 32 Gb RAM DDR2.

In addition, the company, being a part of the «T-Platforms» holding, can provide the resources of a number of supercomputer centers in Russia, with which «T-Services» has partnership relations.

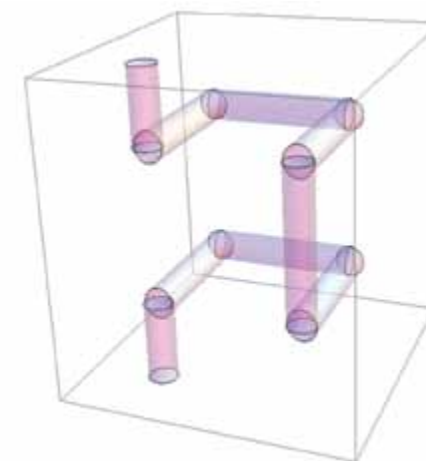
«**NITs SVT LLC**» – small innovative enterprise «Scientific research center of special computing technologies», registered in Nizhny Novgorod in 2010 in order to finish development of software toll platform Cubic(R) and to launch it in the market. The company is a resident of the Nizhny Novgorod Innovation Business Incubator. The project is funded by a private investor, as well as by the fund for promotion of small enterprises in scientific-technical sphere.

The company participates in joint research projects in different fields of physics, biology and medicine, which allows to work closely with

Distribution of tension vector field in biological tissue

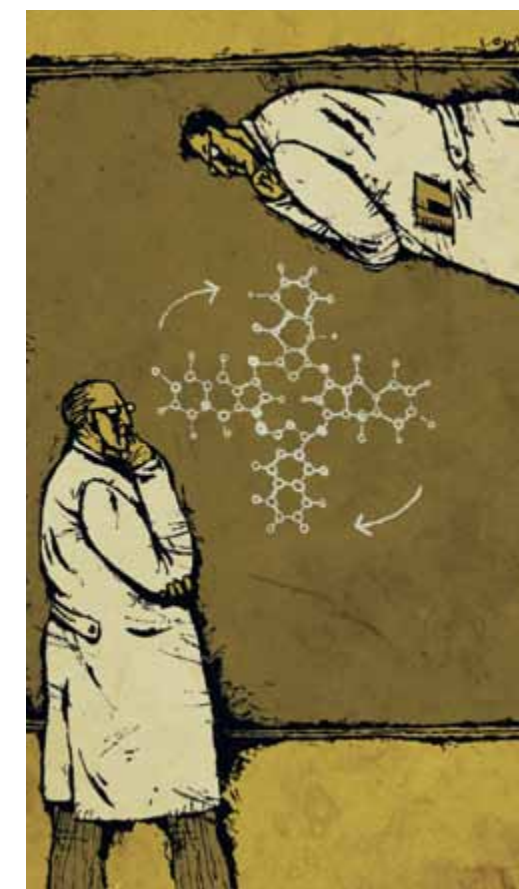


Mathematical models describing both the impact of microwave pulses to living matter and the behavior of the cell mass in the bioreactor, as well as multi-directional hydrodynamic processes in the bioreactor of «artificial liver», are based on the method of cellular automata



Chosen initial configuration of tubes/channels

scientists and applied scientists. Such cooperation helps to more accurately formulate the tasks and eventually to optimize the functionality of the product – Cubic (R) platform, and to refine its components in order to improve consumer options.



About the Supercomputers magazine

Supercomputers is the first and exclusive Russian magazine focused on High-Performance Computing topics, published quarterly by «SCR-Media» Ltd. in Moscow, Russia. The mission of the magazine is to promote effective communication between all Russian HPC market players: manufacturers and consumers of HPC solutions, public authorities and scientific organizations. The editorial content meets the highest world standards for the popular scientific publications and is focused on the cutting edge hardware and software solutions, tools and applications, networking, middleware, storage systems, HPC centers and many more.

The Internet version of the magazine is integrated with a prestigious rating of the TOP50 most powerful supercomputers in the CIS (www.supercomputers.ru), the official analog of TOP500 supercomputers list. Supercomputers.ru is the most comprehensive source in Russia of actual and trustworthy information about the worldwide supercomputing market.

Supercomputers magazine is a member and official publisher of the Russian Supercomputing Consortium of Universities (www.hpc-russia.ru), the main professional association of Russian HPC industry.



"I am in the HPC scene since more than 30 years, starting the ISC events in 1986 and the TOP500 project in 1993 together with my colleagues Erich Strohmaier and Jack Dongarra. In all these years I became acquainted with many supercomputer magazines, especially published in the US. Therefore it was good news to see the 'Supercomputers' magazine from Russia appeared on the HPC scene.

Of course I was delighted to detect my portrait on the cover page of the latest issue. I have met the guys at different ISC events and I can say they work as a professional team and additionally they are very pleasant colleagues."

Prof. Dr. Hans W. Meuer
ISC General Chair
TOP500 Founding Editor
Prometheus General Manager

