*The Second International Conference "Problems of Cybernetics and Informatics"*
*September 10-12, 2008, Baku, Azerbaijan. The plenary paper*
www.pci2008.science.az/1/01.pdf

# SKIF SUPERCOMPUTERS AND THEIR APPLICATIONS

**Sergey Ablameyko[1], Sergey Abramov[2], Vladimir Anishchanka[1], Nikolay Paramonov[1], Oleg Tchij[1]**

[1]United Institute of Informatics Problems National Academy of Sciences of Belarus
[2]The Program Systems Institute of the Russian Academy of Sciences
*abl@newman.bas-net.by*, *otchij@newman.bas-net.by*

## 1. Introduction

Distributed computer systems with mass parallel architecture have become the main line for modern HPC technologies development. Thus, creation of trusted high-performance media for mobile and scalable applications is the current necessity both in science and industry.

Development of supercomputer technologies in any country can be determined by the necessity to further develop and implement latest high-end information technologies, aimed at complex problem solution in mechanical engineering, bio-technologies, land survey & exploration, environment protection, transport and communications, governmental, commercial and other applications. High-performance systems and high-end software products are also widely used in such applications as databases, data computing systems, remote control systems, real-time control systems, etc. In this concern, SKIF supercomputer family, being developed under joint Belarusian-Russian projects, can be found promising. The projects are principally implemented by UIIP NAN Belarus (Joint Scientific Research Institute of Programming & Computer Technologies of the Belarusian National Academy of Sciences), as Belarusian partner, and IPS RAN (Scientific Research Institute of Programming Systems of the Russian Academy of Sciences), as Russian partner.

## 2. Main results

The most important outcome of SKIF project development at the first stage (2000-2004) has become the creation of SKIF cluster pre-production supercomputers of Generation 1 and 2 with the capacity ranging from tens of billion to several trillion Flops. These clusters were applied for both software routine check and in solving practical problems for Russian and Belarusian enterprises and institutions. When realizing the SKIF project, system software and language facilities have been additionally developed for SKIF pre-production supercomputers of Generation 1 and 2.

SKIF Generation 1 configurations were based on single-core 32-bit CPUs, Fast Ethernet, SCI, Myrinet network solutions and 1U-4U constructive form-factors.

SKIF K-500 Generation 2 was designed in 2003, based on single-core 32-bit CPUs, GB Ethernet and SCI network solutions and 1U constructive form-factors.

SKIF K-1000 Generation 2 was designed in 2004, based on single-core 64-bit CPUs, GB Ethernet and InfiniBand network solutions and 1U constructive form-factors.

SKIF K-500 and SKIF K-1000 supercomputers were included in the corresponding Top500 editions.

Top500 lists are updated regularly. For example, the 26$^{th}$ edition (November 2005) includes 351 systems designed in 2005 (70.20%), 94 systems designed in 2004 (18.80%), 35 systems designed in 2003 (7.00%), 15 systems designed in 2002 (3.00%), 3 systems designed in 2001 and one system designed in 2000 and 1999 each. The statistics clearly demonstrate the promising practicability of engineering decisions made by the Belarusian-Russian SKIF supercomputer project team: SKIF K-1000 supercomputer (98$^{th}$ place in the 24$^{th}$ Top500 Edition of Nov. 2004) of 2.5 TFlops peak performance was included in four successive Top500 Editions. SKIF K-1000 supercomputer efficiency counted 80%.

Presently, the top priority of SKIF project development has become creation of SKIF family new generation supercomputer, based on the national scientific background &

engineering innovations and high-end world experience in computer technologies and taking into account modern science-intensive industry requirements.

For this purpose, the new SKIF-GRID program (2007-2010) provides for pre-production supercomputer design of SKIF family Generation 3 and 4 with most perspective computer elements being integrated, which would ensure optimizing performance (power consumption, capacity, weight & dimensions) for SKIF family supercomputers.

In particular, in 2008 new SKIF K-1000M and SKIF MGU supercomputers were designed under SKIF project. SKIF K-1000M Generation 3 high-performance configuration was developed through SKIF K-1000 Generation 2 TFlop-ranged supercomputer modernization.

SKIF K-1000 supercomputer modernization was generally aimed at its functionality expansion and performance increase due to add-in configuration and dual-core CPU integration, as well as at optimizing the supercomputer power consumption.

As a result of such modernization, SKIF K-1000M peak performance doubled to reach 5068.8 GFlops. The practicability of single-core CPU replacement into dual core CPUs was proved by SKIF K-1000M performance tests, demonstrating 1.86 times increased effective performance at ~15% reduction of power consumption.

SKIF MGU Generation 4 supercomputer having the peak performance of 60 TFlops was designed by integrating modern quad-core CPUs and the currently top-performance InfiniBand modification, based on blade-servers in chassis-type form-factors. The SKIF MGU principal designer – T-Platformy Russian Company.

### 3. Supercomputer architecture

Supercomputers family "SKIF" include a number of software compatible supercomputer configurations of extensive range of performance – up to trillions of operations per second. Supercomputer models family "SKIF" creation is based in concept on scalable cluster architecture, realizable on classic clusters of computational nodes, built on general – purpose components (standard microprocessors, memory modules, hard disks and motherboards, including SMP – supporting) [1]. The SKIF cluster architecture is on fig.1.

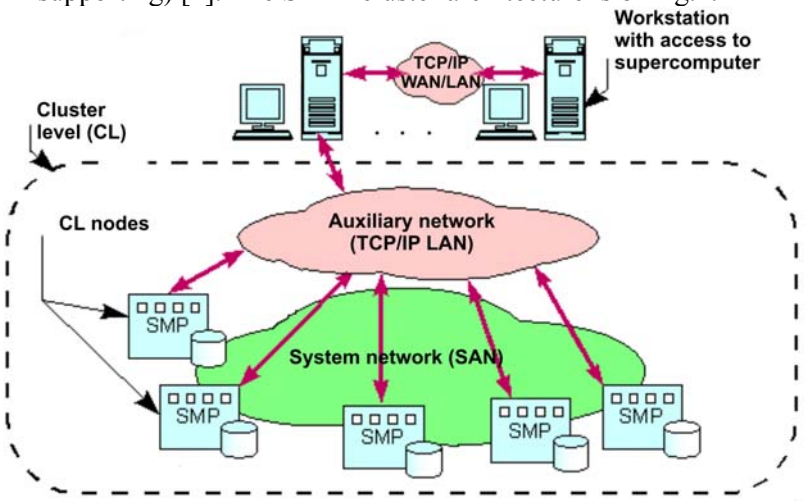


Fig.1. The SKIF cluster architecture

Cluster architecture level is a tightly associated network (cluster) of computational nodes running under OS Linux.

Organization of application task parallel execution is implemented by special systems for supporting parallel computations, that provide effective paralleling various classes applications (usually, of explicitly parallelable tasks): MPI, PVM, Norma, DVM, etc.

Cluster architecture is open and scalable, i.e. it doesn't limit a rigid constraints of a hardware-software platform of cluster nodes, network topology, configuration and supercomputer performance range.

Classical cluster structure with standard hardware components (CPU: Pentium Xeon, Opteron, Itanuim) using different solutions for system network (Gigabit Ethernet, SCI, Myrinet, InfiniBand, etc). Network technology for Gigabit Ethernet, Myrinet and Infiniband is based on ordinary switches communications.

For client data manipulation there are ready solutions which allow to work efficiently with cluster using ordinary PC on almost any platforms. It includes tools for running tasks and retrieve data back. Disk storage of cluster is an ordinary network disk from client point of view. For example, special solutions allow to use CDMA2000 mobile phone connections to work with engineering packages from almost any places in Belarus.

Separate clusters can be integrated into single cluster configuration – higher-level cluster or meta-cluster. Meta-cluster approach allows to create distributed meta-cluster configurations basing on local or global data communications networks. Such meta-cluster configuration was made in the framework of the program "SKIF" between UIIP National Academy of Sciences and PSI RAS to create program "SKIF" informational space.

Meta-cluster configurations-the big step to Grid technologies. That results are the base for next generation of scientific exploration which requires intensive computation and analysis of shared large-scale databases, from hundreds of TeraBytes to PetaBytes, across widely distributed scientific virtual communities. The Grid refers to an infrastructure that enables the integrated, collaborative use of high-end computers, networks, databases, and scientific instruments owned and managed by multiple organizations. Grid applications often involve large amounts of data and/or computing and often require secure resource sharing across organizational boundaries, and are thus not easily handled by today's Internet and Web infrastructures. Grid technologies will be the subject of the new Belorussian-Russian Joint Program "SKIF-Grid".

## 4. Supercomputer System Software

Reference to the aforementioned decisions, system software for cluster supercomputers should be selected to obtain the following major software elements:

1) Cluster operating system;
2) System network software;
3) Parallel program development tools;
4) Batch job processing system.

As noted in preceding section, Linux operating system has been selected as a basic operating system for the cluster solutions. Linux distribution includes various software modules as may be required for proper system operation, administration and program development. Linux distribution may have many parameter-related differences. This report would deal only with those of particular importance for cluster system. For corporate purposes, we can differentiate two manufacturers of this sector – RedHat and Novell.

**MPI** (Message Passing Interface) is considered the prevalent parallel program development tool for distributed memory computer devices. MPI technology presupposes each parallel program to be represented as a set of simultaneous processes, which interact by data exchange through network facilities. At the same time inter-process data exchange should be effected irrespective of CPUs engaged in such interrelated processes. In this respect, MPI is an inter-processor data exchange operation library dealing with different programming languages (C, C++, Fortran77, Fortran90). The MPI project principal is Aragon National Laboratory (U.S.A.).

The major benefit of the message passing standard is its mobility and operational simplicity. Such benefit is even more obvious with distributed memory network environment, which engages high-level procedures and/or abstractions to prevail over message passing

procedures. Moreover, the standard supplies manufacturers with the specific set of procedures to be efficiently dealt with and sometimes to develop hardware options to accommodate the standard.

## 5. State supercomputer multi-access center

UIIP has organized the "State supercomputer multi-access center" (SSMAC)**,** which includes the computational capabilities of cluster supercomputer SKIF K-1000M.

The high speed approach to the resources connected to telecommunication network of the National academy of Sciences of Belarus BAS-NET from scientific networks of Russia is organized through pan European scientific network GEANT. The connection to GEANT is on the base of fiber-optic transmission system with 155 Mb/sec.

SSMAC is connected to BAS-NET with fiber-optic transmission system with 100 Mb/sec. The network interaction of remote users with SSMAC is organized by protocol SSL.

*SKIF K-1000M cluster high-performance computer system software.*

The aforementioned system software elements were successfully tested during SKIF K-1000M cluster supercomputer operation:

Linux – Fedora Core 8 basic distribution kit for x86_64 architecture. Fedora Core project is the RedHat Linux successor and, as such, is financially supported by RedHat. Fedora Core Linux' major features are the latest software, always updated upon distribution kit commercialization, all-times accessibility of free on-line updates. This is essential for cluster applications in educational institutions and for those involved in different research projects.

InfiniBand software – OpenFabrics Enterprise Distribution. The software currently used in SKIF K-1000M – OFED Version 1.2.5.

MPI libraries:

1) MVAPICH

This MPI software for InfiniBand high-speed network by Ohio University is one of the most efficient commercialized MPI products. It is licensed by BSD http://mvapich.cse.ohio-state.edu/. SKIF K-1000M implements the updated MVAPICH Version 0.9.9.

2) OpenMPI

This MPI portable free software was developed by different MPI versions merging (FT-MPI by Tennessee University, LA-MPI by Los-Alamos National Laboratory, LAM/MPI by Indiana University and some others). It is developed and supported by a consortium of academic, research and industry partners. It is remarkable for heterogeneous network support, support of different switchboard solutions (Ethernet, IB over IP), compatibility with a range of BPCS planners, absolute conformity to MPI-2 specification, as declared: http://www.open-mpi.org/.
The version currently used in SKIF K-1000M – OpenMPI Version 1.2.5.

3) HP-MPI

This MPI high-performance commercialized realization by HP conforms absolutely to MPI 1.2 specification and applies some functional elements of MPI 2.0.

HP-MPI is designed for simultaneous operation of all supportable network interfaces with top-fast automatic selection option. For cluster applications, HP-MPI is used to operate copyright packages LSDYNA, StarCD and Ansys family products.

Batch processing system**.** SKIF K-1000M cluster supercomputer incorporates free PBS version - Torque (Terascale Open-source Resource and QUEue Manager), developed by Cluster Resources, Inc. on the basis of OpenPBS. This systems possesses a number of improvements, i.e.

- improved scalability (operation in 2500 nodes environment);
- more robust design (additional testing processes introduced);
- advanced planner interface for advanced and more concrete data;
- improved log registrations.

As the job planner, Maui free software system is used. PBS was domestically optimized for improved cluster stability and performance in various GRID environments (Unicore, gLite).

The version currently used in SKIF K-1000M – torque-2.2.1, job planner – Maui 3.2.6p19.

*Licensed applied software of SSMAC.*

Commercial version of **LS-DYNA** v. 970 for high speed non-linear dynamic processes modeling;

Commercial version of **STAR-CD** for hydro and gas- dynamic processes modeling;

Academic license of **Fluent 6.2** for hydro and gas- dynamic processes modeling.

Engineering systems LS-DYNA (www.lstc.com), STAR-CD (www.adapco.com) adapted for multiprocessor structure allow to solve complex and multi dimensional problems in mechanical engineering with application to real industrial objects design. The Center also has the licensed packages **Pro/E, SolidEdge, Inventor** 9 for creation of computer 3D models of the objects and systems**.** Database **Oracle10g** was successfully installed on SKIF K-1000 nodes for financial bank applications. The recourses could be provided for the users for $0.1 - $0.5 hour/processor depending the monthly volume of the calculation and the period of the renting. The debugging of the remote access and the calculation within one week are free.

## 6. Working applications

There are many applied tasks that have been solved by using supercomputers. We shall not write about all of them. Mention only some:

- automatic recognition and selection of objects in video data stream;
- parallel image processing library (PIPL);
- modeling of turbo compressors for supercharging of Minsk Tractor diesel engines;
- modeling a frame of perspective universal tractors "Belarus";
- modeling a carrier constructions of opencast colliery dump trucks of BelAZ lorries;
- calculation of dynamic characteristics of subsoil cultivators and cardan shafts, produced by Grodno company "Belcard".

## 7. Conclusion

Main results of creation of supercomputers family "SKIF" are described in the paper. Applications of supercomputers are considered.

## Literature

1. S.Ablameyko, S.Abramov, U.Anishchanka, N.Paramonov, O.Tchij, Supercomputer configurations SKIF, Minsk, OIPI NAN Belarusi, 2005, 168p.
2. www.top500.org/system/6715
3. www.top500.org/system/7289
4. Performance of Various Computers Using Standard Linear Equations Software, J. Dongarra, Technical Report CS-89-85, University of Tennessee, 1989.