ISSN 0005-1179, Automation and Remote Control, 2014, Vol. 75, No. 10, pp. 1869–1874. © Pleiades Publishing, Ltd., 2014.
Original Russian Text © N.P. Vasilyev, M.M. Rovnyagin, 2013, published in Avtomatizatsiya v Promyshlennosti, 2013, No. 4, pp. 51–54.

AUTOMATION IN INDUSTRY

Hybrid Clusters for Budget Supercomputers and Cloud Computing

N. P. Vasilyev and M. M. Rovnyagin

National Research Nuclear University MEPhI, Moscow, Russia
e-mail: NPVasilyev@mephi.ru

Received February 15, 2013

Abstract—In the paper, we studied a modern trend in computer engineering, namely, computing clusters with hybrid CPU/GPU architecture which are widely used in different spheres of science and technology. A special attention was given to construction of cheap but sufficiently powerful clusters for industrial computations, process simulation, and construction of own cloud computing.

DOI: 10.1134/S0005117914100130

1. INTRODUCTION

Nowadays, computer engineering media are used in all spheres of sciences and industries. Numerical simulation methods solve complex process problems occurring at the stage of designing of engineering media. Computer simulation in aviation helps considerably reduce prices and increase production of new aircraft: many tests which were previously performed only in the air tunnel now are virtually made on models of future planes. Though planes are available, computations of this level require other computer-based systems rather than personal computers.

In the last years, technologies of cloud computing are used more and more often. In the eye of users, this is an access to computing services as an ordinary one: the user has prepared its task, run it "somewhere" in a provider cloud, and received results of computations. In the eye of providers, they need a flexible infrastructure for providing such services to promptly organize a more suitable computing medium for a user's task: operating system, a list of necessary software, etc. Cloud technologies are based on virtualization of resources and computing processes; high requirements to computing systems of cloud infrastructure are a payment for flexibility.

Special computers exceeding by capacity existing computers by a decade were created from the middle of the XX century for solving complex computing systems. Such computing systems are called supercomputers. The objective of this paper is to review modern technologies for high-performance computations; a special attention was given to hybrid (or "heterogenous" as foreign authors prefer to use) computing systems in which computations are made both by ordinary and graphics processing units. Construction of multicomputer systems (clusters) with hybrid architecture will make it possible to obtain a powerful computing center which later can be used both for solving computationally complex tasks and for cloud infrastructure. At the same time, costs for such a system can be rather modest: this cluster can be made based on available office computers united in an ordinary local network.

2. SUPERCOMPUTERS AND CLUSTER SYSTEMS

The following types of architecture for supercomputers were created and were in demand by the beginning of the 1990s:

(1) Single Processor is a supercomputer based on one valuator;

(2) MPP is a system with massive parallelism which is designed based on standard processors. Interconnect and placement media are, as a rule, proprietary developments of supercomputer vendors;

(3) SIMD is a supercomputer based on vector processors;

(4) SMP is a system combining several processors with a general memory.

Beowulf, a multicomputer system, was put into service in summer of 1994 in space and scientific center NASA. The system was designed based on 16 processors 486DX4/100MHz using standard components installed in ordinary blocks and united by an Ethernet network with capacity of 10Mbit/s. The computing complex had high performance but it was cheaper than MPP analogues in which proprietary technologies were dominating. The idea of "a computing cluster" was further developed with occurrence of new technologies for node commutation. Since 1993, twice a year Top500 supercomputers rating (http://top500.org) is published which is based on testing results of supercomputer capacity by LINPACK program. Capacity of supercomputers is measured in FLOPS (FLoating-point Operations Per Second).

Application of cluster systems is firstly based on computing tasks with a high degree of independence of parallel computations; in queueing and virtualization systems (where it is necessary to assign detached resources for solving a concrete task or provide a service to an individual customer); in systems of high reliability in which there is a necessity for duplication of separate nodes and subsystems and keeping them in "hot standby." Hence, there are several types of cluster systems.

High Availability clusters (HA clusters, failover clusters) is a group of computers that have additional nodes providing for performance of the whole complex in whole in case of failing one or several system components. HA clusters have special program media observing processes in a system.

Load balancing clusters use a method for load balancing on several computing nodes by network and program solutions. Thus we can optimally use all cluster resources, increase capacity, minimize a response time, and avoid overload and failures.

Computing clusters. Since the development of the first Beowulf cluster, tasks solved by HPC clusters (High-performance computing clusters) have not changed. It is still urgent to solve complex computing tasks in different spheres of physics, chemistry, bioinformatics, and other sciences. The architecture of computing clusters strongly differs from systems of high availability and load balancing in which parallelism of tasks dominates. If to consider a HPC system, we speak about parallelism of process, their runtime on different nodes, and transfer of messages between them within one big parallel application. Mainly, nodes of a computing cluster are of the same type and have a limited configuration with a minimum number of input/output subsystems and long-time information storage. The latter is expressed in that many nodes are diskless with no graphic adapter, disk storage, etc.

3. HYBRID COMPUTING TECHNOLOGIES AND THEIR APPLICATION

Development of hybrid supercomputer systems (in modern understanding) in many respects is related to the history of evolving of graphics accelerators. First graphics accelerators created for increasing a rate of graphic data display were able only to rasterization (transformation of triangles to pixel arrays). As soon as they are developed, graphics processors attain new functionality, e.g., an ability of vertex processing and processing of counting microprograms for dynamic illumination from several sources.

Most of operations with graphics are quite amenable to paralleling. This fact determines the existing architecture of graphics processor in which there is a great number of functionally independent processor cores with a small volume of cache memory and primitive structure of pipeline.

Nowadays, a great part of supercomputers holding high positions in TOP500 have a hybrid architecture. For accelerators, processors NVIDIA 2050, NVIDIA 2070, AMD FirePro S10000, etc. are used. The first place in the 40th anniversary rating of TOP500 (November of 2012) was taken by Titan Cray XK7 in the National Laboratory of Oak Ridge. Titan has a hybrid MPP architecture (designed based on processors AMD Opteron 6274 and accelerators Nvidia Tesla K20X), the total number of cores is 560 640 pieces including 299 008 CPU cores and 261 632 (SMX) GPU cores. With the capacity equal to 17.59 PFLOPS, the supercomputer has a fine index of capacity/working power attaining the value of 2.14 TFLOPS/KW. Its Japanese competitor, supercomputer of a classic non-hybrid architecture (K computer) by Fujitsu, this index is only 0.83 TFLOPS/KW with commensurable capacity.

Thus the technology of hybrid supercomputer computations has received a wide application. More powerful solutions have an MPP architecture and use proprietary interfaces as an interconnect. Supercomputers from the intermediate and high sectors by capacity have, usually, a cluster architecture and use standard technologies of machine-to-machine communication such as 100-gigabit Ethernet and InfiniBand.

Application of hybrid computing clusters in industries need a special attention. The electronic resource http://www.nvidia.ru/object/cuda_app_tesla_ru.html has a list of both different scientific-technical and industrial trends and concrete program products oriented on hybrid computations. These are products such as AutoCAD, Matlab, Simulink, Mathematica, SEMCAD, Gauda, etc.; a list of these applications widely used in industries is great.

Application of hybrid clusters is also important for creating cloud infrastructures. The fact that hybrid technologies provide for a considerably high relative capacity as per kilowatt power as against classic ones, namely, non-hybrid systems, is significant. These issues were studied in [1].

4. HARDWARE AND SOFTWARE FOR HYBRID CLUSTER SOLUTIONS

Modern cluster systems have a wide range of devices and technologies used for their construction. Some of them have gained a great application.

Solutions based on Ethernet and InfiniBand are the most popular technologies for internode connections. The capacity of Ethernet and InfiniBand is very high. InfiniBand is easy for scaling; it is possible to unite basic bi-directional buses in groups (1x, 4x, 12x). There are several modes of operation: Single Data Rate (SDR), Double Data Rate (DDR), Quad Data Rate (QDR), fourteen data rate (FDR), and enhanced data rate (EDR) with basic rates of bi-directional buses 2.5, 5, 10, 14, 25 Gbit/s respectively. One of the main advantages of InfiniBand is application of a simple differential scheme for exchanging signals on a physical level. The length of InfiniBand lines is limited by 17 m. Thanks to poor power consumption, it is possible to create single-chip integral InfiniBand commutators with a great number (> 32 pcs.) of integral transceivers. InfiniBand has standard approaches to organization of RDMA exchanges, IPoIB (IP over InfiniBand), SRP (SCSI RDMA Protocol), and other protocols.

In 2010 a new IEEE Std 802.3ba-2010 appeared describing an advanced technology for data transmission by 100-gigabit Ethernet. Keeping the total continuity within the line of standards 802.3 (a framing format of Ethernet 802.3, minimum and maximum sizes of FrameSize format are preserved), a physical level (PHY) was considerably overworked. Largest vendors (Mellanox, Reflex Photonics, Sumitomo Electric Industries, OpNext) have presented their solutions for creating commutation supercomputer subsystems using CFP modules for multimode and singlemode optical fibers for the physical level. Solutions with optical switch fabric for creating motherboards are even more promising in terms of cost, reliability, and power features. Apparatuses of standard IEEE 802.3ab (Gigabit Ethernet) are still dominating in the market of available interconnect media. Devices designed according to this technology have considerable capacity and low price. However, high latency ($> 100\ \mu$s) aggravates their use in high-performance solutions.

As for computation accelerators, devices of companies NVIDIA, ATI and PLD matrices by Xilinx and Altera are often used.

FPGA can perform functions of any device if to load its logical description into them. The loading process takes several seconds and can be performed an unlimited number of times. A concept of “one task processor” has become popular with researchers and engineers in supercomputer computations (http://school-2010.hpc-russia.ru) which is in providing nodes of a computing cluster with boards with FPGA microarrays. This approach has several problems: FPGA are lower by a decade than “rigid” logic in a rate of computations and a gain from paralleling of tasks on FPGA accelerators should be as minimum to compensate this delay; no acceptable code transformation media for the program model for presenting computations in an array-based logical description of an array; a high cost of FPGA.

ATI company proposes its technology for parallel computations AMD FireStream. The first devices supporting this technology were Radeon X1900 video cards. Further the technology has been improving. A cross-platform language of OpenCL is used as a main language for developing parallel modules [2]. The cross-platform portability of OpenCL solution is achieved by giving a programmer abstract platform models, memory, runtime, and programming. Writing and checking of programs is ensured by ATI APP Software Development Kit. A main disadvantage of ATI is complexity in writing of program modules on OpenCL as the language has a low level.

Technology NVIDIA CUDA has become very popular with developers of high-performance cluster systems. A program model CUDA (Compute Unified Device Architecture) makes it possible to introduce a description of parallelism and memory hierarchies into a programming language. The technology has a hardware support for parallel computations, a cross-platform compiler, and a runtime system.

To simplify “navigation” on the whole set of CUDA devices, NVIDIA introduced a notion of compute capability expressed by a pair of numbers separated by a period: major.minor. Here major denotes a global architectural version and minor is for small changes. The first CUDA valuators of series 8800 had compute capability 1.0. By the moment this paper was written, accelerators with Kepler (Nvidia Tesla K20X) architecture had compute capability 3.5. The second generation of GPU with architecture Kepler (K20 and K20X) has peak capacity for computing single-precision with floating point 3.52 and 3.95 TFLOPS (http://www.nvidia.com).

Apart from compilers of high-level languages and architecture media for parallel computations, each computing node of the cluster should have libraries for machine-to-machine communication. The most popular technology for programming internode communication is MPI (Message Passing Interface). An interface for communication in MPI [3] is only for information exchange in computing systems with a distributed memory. According to the MPI concept, to solve a task, one parallel program is developed which is run simultaneously on all processors of the supercomputer system. As per the standard (the last one was from September, 2012), MPI is program devices with a set of data transmit operations organized as a library for program modules for C and Fortran languages. The library is based on message communication functions between processors. Transmit operation can be both point-to-point and collective. The standard is implemented on a set of free of charge and commercial libraries: MPICH, Open MPI, LAM/MPI, Intel MPI, HP-MPI, etc.

A choice of system software (SS) can be a difficult task as in the market there are a lot of different distribution disks, compilers, and libraries from different vendors. There are also different distribution disks specially collected for using in cluster systems.

Possible variants of SS:

- OS Windows, a set of libraries for computations, compilers, and drivers of NVIDIA;
- an OS similar to UNIX, e. g., one of distribution disks of Linux and an appropriate number for compilation and run of programs;
- a specialized distribution disk for creating a high-performance cluster solution, e. g., CLIC or Rocks. In this case, the checking process will be partially automated;
- a LiveCD distribution disk for checking a dynamic cluster (PelicanHPC GNU Linux). Here the OS on consoles is run from a disk and is adjusted automatically and user's programs and files are stored in a virtual catalogue in RAM. Files are loaded through a network.

5. DESIGN OF A BUDGET HYBRID CLUSTER

The following configuration can be used for designing a low budget hybrid computing cluster. As for communication media, to use a compiler and adapter of IEEE 802.3ab (Gigabit Ethernet) standard. To design computing nodes based on Intel platform (for using optimized libraries, e.g., Intel Math Library). A central processing unit should not be multicore (≥ 6 cores) as in a hybrid computing cluster the CPU performs a role of task planner. Now the best indices as per capacity/price belong to four-core processors of Ivy Bridge and Sandy Bridge (Socket 1155) architectures. A motherboard chosen as per the processor should be on a qualitative hardware components and provide for stable power to the processor. Availability of additional abilities when the processor has run and switching of additional devices is not compulsory in this case and will only lead to increasing the price. The volume of operative memory should the largest as is available according to the budget (≥ 4 Gb). As for a computation accelerator, we can use video adapters in a user's sector of a market as they are cheaper than professional solutions like Tesla. It often makes sense to choose graphics adapters from a previous model range as their prices are considerably lower since the demand is oriented to new models. NVIDIA marks high-performance adapters with a GTX prefix; it is necessary to pay attention to them. For a controlling node (console) of the cluster, it is not necessary to have a graphics CUDA valuator and a processor more powerful than for the nodes. The console should be provided with an additional network adapter for accessing from an outer network.

If each node is equipped with two network adapters, we can organize a physical network separation for MPI and NFS message transmit. For this we need a usual or controlled switch with a double number of ports and ability to create VLAN. After configuration the computing network like this, we can move to diskless nodes of the cluster and arrange a total storage on consoles or on a separate node.

After setting an OS, drivers of video cards, network equipment, and libraries, we can adjust applied software. Apart from the program packages, there is enough of freely distributed software. As a result, we can get an own low budget supercomputer.

6. CONCLUSION

Hybrid computing clusters deserve your attention as they are an inexpensive method for creating own supercomputers. This cluster can be organized based on available computers equipped with modern discrete video cards and united in a local network. All modern OSs including freely distributed have a support for hybrid clusters. A support of virtualization on hardware (built-in modern microprocessors) and software (supported by modern OSs) levels makes it possible to use these clusters as infrastructure for cloud computing. A wide choice of applied software (both proprietary and freely distributed) allows solving tasks of process designing and simulation.

REFERENCES

1. Chernyak, L., A Cloud, *SuperComputers*, 2010, vol. 2, no. 2.
2. Linev, A.V., Bogolepov, D.K., and Bastrakov, S.I., *Tekhnologii parallel'nogo programmirovaniya dlya protsessorov novykh arkhitektur* (Parallel Programming Technologies for Processors of New Architectures), Moscow: Mosk. Gos. Univ., 2010.
3. Gergel, V.P., *Sovremennye yazyki i tekhnologii parallel'nogo programmirovaniya* (Modern Languages and Technologies of Parallel Programming), Moscow: Mosk. Gos. Univ., 2012.