Organization of cluster systems based on the PCI Express switching environment

N S Vasyaeva and E S Vasyaeva
Volga State University of Technology, Lenin Sq., b. 3, Republic of Mari El, Yoshkar-Ola, 424000, Russia

E-mail: vasyaevans@volgatech.net

Abstract. This article is devoted to the organization of computing cluster systems that have a limit on the maximum distance between computing nodes and data storage nodes. Known options of building clusters based on the PCI Express switching environment are considered as the most.

1. Introduction
Modern methods of increasing the computational efficiency of cluster systems lead to the need to improve the switching structures implemented in the cluster. This due to the fact that the switching structure is the bottleneck of the entire cluster system [1, 2], since with increasing node performance, the need for fast data transfer between nodes increases. This leads to downtime and a decrease in overall cluster performance.

There are many approaches to designing a switching system for clusters: tree-form, cubic, hypercubic, toroidal, and other structures can be used [1, 2]. The choice of the cluster switching network is primarily determined by the characteristics of the transmitted traffic.

In some cases, when intensive data exchange between computing nodes is not required, the use of widespread Ethernet family of technologies is sufficient. The cluster RHIC Computing Facility (Brookhaven National Laboratory, USA) based on Ethernet technology was built. In other cases, when the need arises to make a large number of inter-node communications, low-latency networks with high throughput rates are used, for example, InfiniBand and Fiber Channel networks. In particular, on the InfiniBand FDR network, an inter-node connection of the MVS-10 П supercomputer of the Interdepartmental Supercomputer Center of the Russian Academy of Sciences was built. The Gbit Ethernet network is used as a service network in this project.

Own switching structures and protocols are created for large projects that occupy leading positions in the TOP500 list of supercomputers. The high-speed Angara network (NITSEVT JSC) [3, 4] is an interesting solution. A multidimensional torus (4D-tor) topology is implemented for communication between nodes in this network and the interaction of the network adapter with the computing node is carried out through the PCI Express Gen2 x16 interface.

There are also standards, for example, PCI Express, which, in cases of creating a computing cluster with a small territorial fragmentation of nodes (up to 100 meters), are not inferior to high-performance solutions, such as InfiniBand, and have the advantage of lower financial deployment costs (cost of one InfiniBand switch ports range from $ 300 to $ 1,000, and PCI Express ports from $ 170 to $ 300). In addition, InfiniBand network capabilities are redundant, for some cluster systems with topological
limitations. In this case, an acceptable solution is a switching network based on cable implementation of PCI Express technology.

2. Approaches to the construction of cluster systems based on the PCI Express standard

At the heart of any PCI Express topology is a root complex and switch, applied to external cable switching allows you to create geographically separate independent clusters (due to restrictions on the length of the cable system), or serves to unite a team of computers into microclusters, which in turn are combined to macro clusters using the InfiniBand standard or other high-performance protocols. The maximum length of a PCI Express cable for on-board connections is 15-30 meters, depending on the type of cable.

The construction of a computer network or cluster using PCI Express technology has its own peculiarities in that either the switch or the node must use special nontransparent bridge or ports using the Non-transparent Bridging technology [5]. This technology is designed to distinguish between address spaces of upstream and downstream devices connected to the network via the PCI Express bus, that is, to isolate different domains of the PCI Express bus hierarchy physically belonging to different devices. The topology of the PCI Express system bus is similar to the standard network topology built on switches, moreover, the network traffic is switched and the switches use efficient flow arbitration algorithms. At the same time, the transmitted PCI Express packets contain address information that allows data to be transferred outside the boundaries of one local system and to bind hierarchy domains located in different address spaces (figure 1).

![Figure 1. PCI Express technology sample, that combines two independent address domains.](image)

2.1. Communication environment MVS-Express

There are several approaches to creating cluster systems based on the PCI Express standard. One of them was developed in Russia in the Research Institute "Quantum" and IPM them. M.V. Keldysh RAS for the computing system "MVSExpress" [6, 7]. The MVS-Express communication network builded using the direct switching technology of PCI Express packets and implements at the hardware level a common field of large-capacity memory and external devices for all nodes in the cluster [8]. Its generalized structure is shown in figure 2.

At the same time, the levels of speeds and delays in transmitting data in such a network are comparable to similar indicators for data transfer within a single computing node.
Figure 2. The structure of the communication environment "MVS Express" CN - compute node, IB-interface board, TB - transparent bridge, NB - nontransparent bridge.

The MVS-Express communication environment consists of one master node connected to the switch through a transparent port, and many slave nodes connected through nontransparent ports. The master node is responsible for the distribution of the shared address space. The external switch is engaged in addressing the transfer of packets between computing nodes. Each computing node, which is a server with a PCI Express bus on the motherboard, has a specialized interface board for connecting to the PCI Express cable system, which is configured in the nontransparent port mode (except for the host node). The system as a whole has high performance indicators:

- delay for single write to remote RAM: 0.2-0.4 microseconds;
- transfer rate from 600 MB / s to 1.5 GB / s depending on the channel width;
- the length of the data packet on which the peak speed is reached is several tens of bytes (the average in other systems is several kilobytes).

The specialty of the "MVS Express" system is the use of one of the computing nodes as a control node, which generates heterogeneity of cluster nodes, changes the hardware and software implementation of the head node compared to the slaves nodes.

2.2. Computing cluster based on PCI Express bus from IDT

Another way to create a computing cluster topology based on the PCI Express bus is proposed by the American company IDT (Peter Z. Onufryk, Cesar A. Talledo) [9]. In this solution stated that many computers, which have a processor, RAM and PCI Express root complex in their composition, can interact with each other through a nontransparent switching mechanism (figure 3).

This approach involves multiple address translation when transferring from one node to another, because the packet is forced to pass through several nontransparent bridges. However, unlike the MVS Express system, this system consists of equivalent computing nodes, and the entire load of switching traffic and displaying the common address space lies on the external switch. This is a distinctive feature of the organization of the PCI Express switch proposed by IDT.
2.3. **Computing cluster based on PCI Express bus from IDT**

LSI [10] proposed another implementation of the computing cluster based on the PCI Express standard. This solution uses a modified PCIe architecture, which is designed for interoperability between several independent PCI Express hierarchies (figure 4). The system consists of four nodes with different address domains PCI Express “Local PCI Express hierarchy 1-4”. Each of the four nodes has its own “PCI Express Root Complex”, which distributes the address space in the local bus domain, an internal PCIe switch that connects all local devices on the bus, as well as an external port (cluster port), which is controlled using configuration registers and serves to join the node into a cluster. Configuration registers contain BAR address registers, which serve as “address windows” for communication with other cluster nodes. Thus, each node allocates a certain range of addresses, which is set in the cluster port to be able to access the node.
Figure 4. Cluster structure on PCI Express switching environment proposed by LSI Corporation.

Each of the cluster ports connected to an external PCI Express switch over a dedicated bus. From the side of the external switch, each base address of the local node is associated with its virtual bus number. The switch routes the transactions in accordance with the high-order bits of the memory address, which are used to identify a specific cluster port.

This method of connecting nodes to the cluster using the PCI Express protocol implies the use of direct address translation only, which has a clear drawback - the number of nodes connected to the cluster by one switch is limited by the number of BAR address registers in each port, and there are only six of them according to the PCI Express standard.

3. Conclusion
Almost any switching system where the use of switching equipment is provided, various methods of translating the addresses of computing nodes that introduce delays in the passage of traffic are used, which is critical for large numbers of forwarding operations between nodes and especially for frequent exchanges of small pieces of data. The most preferred option is to use switches with nontransparent ports, because they can broadcast in two modes: fast live broadcast is used with a small number (up to 6) nodes served by the switch, table translation is more labor-intensive and is used with a larger number of connected nodes.

Switches that translate addresses are used in many topologies of computing clusters, linking as separate segments with different topologies (with a mixed topology), or as a forming link in topologies such as “star”, “thick tree” and others. Thus, an urgent task at present is the development of new methods for translating the addresses of nodes of a computing cluster.

References
[1] Karavay M F 2011 Distributed full switch as an “ideal” system network for multiprocessor computing systems Management of large systems 34 92-116
[2] Podlazov V S 2006 High-dimensional single-stage switches for multiprocessor and multi-
machine computing systems Problems of management 6 19-24

[3] Zhabin I A 2014 The first generation of high-speed communication network "Angara" High technology 1 21-7

[4] Korzh A A, Macagon D V, Zhabin I A and Syromyatnikov E L 2010 Domestic communication network 3D tor with support for globally addressable memory for supercomputers transetaflops performance level Proc. Int. Sci. Conf. Parallel Computing Technologies, PaVT'2010 (Chelyabinsk, Russia: South Ural State University) pp 227-37

[5] Kwok K Non-transparent Bridging with IDT 89HPES32NT24G2 PCI Express® NTB Switch

[6] Latsis A O, Davydov A A, Lutsky A E, Smolyanov Yu P, Chetverushkin B N and Shilnikov E V 2010 Multiprocessor computing system of hybrid architecture MVS-Express Reports of the Academy of Sciences 434(4) 459-63

[7] Levin V K Chetverushkin B N, Elizarov G S, Gorbunov V S, Lacis A O and Korneev V V 2014 Switching network MVS-Express Information Technologies and Computing Systems 1 10-24

[8] Chetverushkin B N, Smolyanov Yu P and Lacis A O 2012 Patent 2461055 RF IPC G06F13/00 Cluster system with direct switching of channels 25 p 13

[9] Peter Z O and Cesar A T 2013 Patent US8429325 B1 PCI Express switch and method for multiport non-transparent switching

[10] Kloepner J R, Gates D E and Stubbs R E 2008 Patent WO2008/105933A2 US G06F13/36 Apparatus and methods for clustering multiple independent pci express hierarchies