Development of functional fault-tolerant system

K Zhigalov\textsuperscript{1,2}, S Yu Kuznetsova\textsuperscript{3} and M V Sygotina\textsuperscript{4}

\textsuperscript{1}V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia
\textsuperscript{2}Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Russia
\textsuperscript{3}Irkutsk National Research Technical University, 83, Lermontov Str., Irkutsk, 664074, Russia
\textsuperscript{4}Bratsk State University, 40, Makarenko Str., Bratsk, Russia

kshakalov@mail.ru

Abstract. Before you begin to develop a fault-tolerant system, you should specify the system requirements. Besides high fault tolerance and performance, the computing system shall be highly available. Such systems include NUMA systems with non-uniform memory access, SMP systems with symmetrical multiprocessing and MPP mass-parallel systems. The ultimate choice of architecture largely depends on whether maximum performance, scalability, or fault tolerance and availability are needed.

1. Introduction

Reliability and fault tolerance of the technical objects is the essential aspect of the system science. All types of the manufacturing \cite{1}, mining \cite{2}, telecommunication \cite{3} or information \cite{4} systems should match their own criteria of reliability. Achieving this required level of reliability requires special measures: hardware \cite{2, 5}, organzational \cite{2, 6} or software \cite{3, 4}. This paper is dedicated to the discussion on ensuring fault-tolerance in computer cluster. The test bench and series of the experiments \cite{7} are used to assess performance, hardware resource utilization and reliability.

Among different versions of systems, a cluster system seems the most relevant for the task. The cluster system represents two or more nodes that are interconnected by a high-speed data link, and shares resources for storing jointly processed data that work together to perform a common task and appear to the user as a single system. Each cluster node runs its own operating system. The cluster system can reach 99.999\% availability. Despite the high cost of this solution compared to a standby server, fault tolerance and availability often become important in choosing the system design scheme.

Figure 1 schematically shows a two-node cluster system. \cite{8}

A cluster node can be a single-processor computer or a multiprocessor computer, and clusters can be completely different in configuration. Cluster nodes are connected by high-speed communication channels using Ethenet, Fibre Channel or firm-specific technologies. Across cluster connections, nodes do not only exchange information, but also monitor the availability of the neighboring nodes.

2. Description of a fault-tolerant system

Cluster systems shall have high availability (otherwise alert level) and be insensitive to failures of one of the nodes (both hardware and software). In case of failure of one of the nodes, the entire load shall be automatically transferred to other nodes of the system \cite{9}. Once a cluster system node fault is
corrected, all services and applications can be relocated back to it.

Figure 1. Two-node cluster scheme

Depending on the type of backup nodes of the cluster system, there are three types of failover clusters (Figure 2).

Figure 2. Types of cluster systems

In a system with cold standby of nodes, the active nodes perform the work, while the passive nodes wait for failure of one of the active nodes and, if necessary, are put into operation.

In a hot standby system, all nodes are included in the operation. In case of the failure of one of the cluster nodes, all load is distributed among the remaining nodes. Such a cluster is also called the request distribution support cluster.

The cluster system with modular redundancy is rarely used in practice because of its high cost and only when a simple system is not acceptable. This system guarantees that the nodes perform the same
work and the result will always be the same, or failure of one node will not affect further work. An example of such a system is RAID 1 disk array where all data is written into two disks in parallel (mirrored) and the failure of one disk will not result in the information loss.

Another classification of cluster systems may be the availability or absence of shared resources. A cluster with shared disk arrays has access to a shared disk space and must have a distributed lock manager to manage the queue disk write and prevent conflicts. In such a system, local disks are mainly used for the initial boot of the operating system.

3. Used equipment
The test bench will be HP ProLiant DL180 Gen9 servers with the following characteristics:

- 2 Intel® Xeon® E5-2600 v3 processors;
- 64 gigabytes of RAM DDR4 SmartMemory;
- Dynamic Smart Array B140i controller
- 4 HP SAS HDD 2.5 hard drives with 1.2 gigabytes combined in RAID 10.

The nodes are running Windows Server 2012 R2 Datacenter, which allows them creating virtual machines based on UEFI (Unified Extensible Firmware Interface) and deploying virtual machines twice as fast as 2008 version.

The Windows Server family of operating systems is a server-centric operating system that focuses on cloud computing and virtualization. Compared to an earlier version, it has, for example, updated Hyper-V version, IP address management role, new version of Windows Task Manager, new file system. Besides, unlike its predecessor, Windows Server 2012 can switch between “Server Core” and “Server with GUI” without being completely reset. The Server Core is only the command line interface option, which is now recommended configuration. There is also a third setting option that allows running some GUI elements, such as MMC and Server Manager, but without the usual desktop programs, and programs such as the Explorer.

The Datacenter license allows deploying an unlimited number of virtual machines and creating failover clusters that support dynamic failover in the event of failure, and scaling the nodes according to tasks.

Also, for high availability and fault tolerance, the hard drives are combined into RAID 10 (Figure 8). RAID works by placing data on multiple disks and allows I/O operations to be performed symmetrically thus improving performance.

RAID 10 (RAID 1+0) is RAID 1 (mirroring) and RAID 0 (improved disk array) that provides better performance than RAID 1, but with much higher performance. In RAID 10, data is mirrored and the mirrors are interleaved.

In such a scheme, in addition to combining two physical disks into one logical (to increase the volume), the data is duplicated to the second exactly the same disk. This redundancy scheme avoids data loss when a single copy fails. The second copy of RAID 1 of the array will continue to function, and if the failed disk is replaced from the faulty mirror, the second (mirror) copy will be fully restored to the double.

4. Shared resources
A failover cluster system requires a shared resource, such as disk space or data storage system. The nodes read information from it, process, and record the result of performed work. The shared resource will be an HP StorageWorks P2000 G3 disk array with the characteristics shown in Table 1.
Table 1. Disk array characteristics

| System software | HP StorageWorks P2000 G3 |
|-----------------|--------------------------|
| number of controllers | 1 or 2 (Active/Active) |
| RAID levels | RAID level 0, 1, 3, 5, 6, 10, 50 |
| cache memory | 2GB cache memory for every controller (automatic cache memory replication into non-volatile memory in case of power failure) |
| hard drive | up to 12 hot-plug 3.5" LFF SATA/SAS HDD |
| extension | extended by linking to up to 7 disc enclosures P2000 12-drive Enclosure (12 x 3.5" LFF HDD) or up to 5 disc enclosures D2700 25-drive Enclosure (25 x 2.5" SFF HDD) |
| maximum capacity | 288TB (96 x 3.5" LFF HDD) |
| maximum number of hard drives | 137 (12 x 3.5" LFF + 125 x 2.5" SFF HDD) |
| power sources | 595W fault-tolerant (1+1) power supply system with hot standby of power sources |
| management | HP Storage Management Utility (SMU), USB (Command Line Interface), 10/100 Ethernet (WEB Interface, SNMP, SMI-S, SSL, SSH, SMTP, FTP, HTTP, Telnet) |
| system software | offline backup SnapShot, VolumeCopy and RemoteSnap |

There are three main types of data storage connectivity:

- DAS (Direct Attached Storage) – direct connectivity to a server;
- NAS (Network Attached Storage) – network data storage system;
- SAN (Storage Area Network) – data storage network.

DAS is a traditional way to connect directly to one or more servers via high-speed interface. Among the above storage management methods, DAS is the fastest and recommended for read/write systems. The connection is based on SAS (Serial Attached SCSI) technology, which is a serial interface with speeds of up to 24 Gb/s and a Fibre Channel.

NAS is a file-level file storage server connected to a computer network that provides access to data for a group of clients. It is usually a budget choice and allows installing only a few hard drives, usually up to 4.

SAN is an architectural solution for connecting external storage devices such as disk arrays, tape libraries, optical drives to servers so that the operating system recognizes the connected resources as local. Although the cost and complexity of such systems are constantly falling, the storage networks remain rare outside large enterprises. SANs are characterized by the so-called network block devices (usually through Fibre Channel, iSCSI, or AoE protocols), while NAS network storage is aimed at providing access to data stored on their file system through a network file system (such as NFS, SMB/CIFS, or Apple Filing Protocol).

Due to its flexibility in SAN, like LAN, it is possible to create different topologies and technologies, but it is through the Fibre Channel that it has gained popularity.

Fibre Channel (FC) is a high-speed network technology (typically running at speeds of 1, 2, 4, 8, 16, 32, and 128 gigabit per second) that is mainly used to connect data storage to servers and typically runs on fiber optic cables, but can also run on copper cables. Fibre Channel combines the best technologies from both channel and network interfaces:

- High speeds;
- Protocol independence (0-3 levels);
- Long distances;
- Low delays;
- High reliability;
- High scalability;
• Multipoint topology.
  Fibre Channel is an open standard for serial, high-speed interfaces. It allows connecting servers and storage systems up to 50 kilometers or up to 400 megabytes per second (up to 10 kilometers). All this allows building the SAN topology following the same principle as traditional networks, based on switches and hubs, and allows organizing the system without a single point of failure.

5. Internode connections
Internode connections play an important role in building a high-performance and fault-tolerant system. Figure 3 shows some of the main node connection topologies in high-performance computing systems. It should be considered that choosing the type of internode connection will determine the characteristics of the entire system, which will affect its performance and fault tolerance.

![Types of internode connections](image)

Figure 3. Types of internode connections

It should also be understood that in the development of a cluster system, the use of a single physical data medium for internode connectivity and for the local network will not only increase the time of exchange between the nodes, but will also increase the possibility of failure of the entire system. Therefore, this study will use a separate dedicated physical link for internode connection only, built on Gigabit Ethernet technology using a Category 5E UTP cable.

Virtualization is the creation of a virtual (not actual) version of something, including virtual computer hardware platforms, storage devices, and computer network resources. Hardware virtualization or platform virtualization refers to creating a virtual machine that acts like a real computer with an operating system. The software running on these virtual machines is separate from primary hardware resources. For example, a computer running Microsoft Windows may host a virtual machine that is similar to a computer running Ubuntu Linux. Virtualization began in the 1960s as a method of logically dividing the system resources provided by mainframes between different applications. The meaning of the term has since greatly expanded.

Virtualization refers to creating a virtual resource, such as a server, a desktop, an operating system, a file, a storage, or a network.

The primary goal of virtualization is to manage workloads by allocating computing resources to make them more scalable. For decades, virtualization has been part of the IT infrastructure, and today it can be applied to a wide range of system levels, including virtualization at the OS level,
virtualization at the hardware level, and server virtualization.

In hardware virtualization, a host machine is a physical machine on which virtualization takes place, and a guest machine is a virtual machine. The words “host” and “guest” are used to distinguish between software that runs on a physical computer and software that runs on a virtual machine. The software that creates a virtual machine on host hardware is called a hypervisor or a virtual machine manager.

There are three types of virtualization:

- Software virtualization (also known as paravirtualization) – hardware environment is not modeled, but guest programs run in their isolated domains as if they run in a separate system. Guest programs must be specially configured to work in this environment;
- Hardware virtualization – a way to improve overall virtualization efficiency. It includes processors that provide hardware virtualization support and other hardware components that help improve guest performance. Hardware virtualization is used to simulate a full hardware environment or a virtual machine in which an unmodified guest operating system (using the same instruction set as the host computer) is executed in full isolation (Figure 10);
- Virtualization at the OS level (also known as containerization) – refers to an operating system feature in which the core allows multiple isolated instances of user space to exist. Such instances, called containers, virtualization mechanisms, or prisons (prison FreeBSD or chroot jail), may look like real computers in terms of the programs they run.

*Figure 4. Hardware virtualization*
6. Conclusion

Virtualization can be divided into different levels: desktop, server, file, storage, and network. Network function virtualization reduces the number of physical components, such as switches, routers, servers, cables, and hubs, that are needed to create multiple independent networks, and this is particularly popular in the telecommunications industry. When virtual nodes are used to create a failover cluster system, you simply need virtual network consciousness.

To support hardware virtualization, you must enable the appropriate option in the BIOS or UEFI.

References

[1] Seung M, Hyung O Venture J, Joon K & Yong-jin Y 2013 Service reliability improvement in manufacturing and operating systems International Journal of Precision Engineering and Manufacturing 14 1401-1406.

[2] Kozhiev H H, Klyuev R V, Bosikov I I and Youn R B 2017 Analysis of management of mine ventilation networks using simulation models Sustainable Development of Mountain Territories 9 (4) 414–418

[3] Mytsko E, Malchukov A, Kim V, Osokin A, Zoiev I, Ryzova S 2016. Software implementation research of CRC computation algorithms compatible with PKZIP, WINRAR, ETHERNET Advances in Computer Science Research doi: 10.2991/itsmssm-16.2016.16.

[4] Pokorni S 2014 Reliability of Information Systems 2014 (Belgrad: Information Technology School).

[5] Faerman V, Avramchuk V, Luneva E 2014 Prospects of frequency-time correlation analysis for detecting pipeline leaks by acoustic emission method IOP Conference Series Earth and Environmental Science 21(1) 012041.

[6] Sulzer-Azaroff B, Santamaria M 1980 Industrial safety hazard reduction through performance feedback Journal of applied behavior analysis 13 287-95.

[7] Cheremnov A et al 2017 Optimization of the coherence function estimation for multi-core central processing unit IOP Conf. Ser.: Mater. Sci. Eng. 177 012020.

[8] Zhukovskiy Y, Batueva D, Buldysko A, & Shabalov M 2019 Motivation towards energy saving by means of IoT personal energy manager platform. Journal of Physics: Conference Series 1333(6) 062033

[9] Zhukovskiy Y, Malov D 2018 Concept of Smart Cyberspace for Smart Grid Implementation IOP Conf. Series: Journal of Physics 1015(4) 042067

[10] Urazbakhtin R Yu, Yungmeyster D A 2019 The results of studies of the tunneling rescue complex for coal mines, IOP Conf. Ser.: Mater. Sci. Eng. 560 012130