ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-THPHA048

NEW IT-INFRASTRUCTURE OF ACCELERATORS AT BINP

P. B. Cheblakov *, F. A. Emanov, D. Yu. Bolkhovityanov, BINP SB RAS, Novosibirsk, Russia

Abstract

In 2017 the Injection Complex at Budker Institute, Novosibirsk, Russia began to operate for its consumers - colliders VEPP-4 and VEPP-2000. For successful functioning of these installations is very important to ensure a stable operation of their control systems and IT-infrastructure. The given article is about new IT-infrastructures of three accelerators: Injection Complex, VEPP-2000 and VEPP-4. IT-infrastructure for accelerators consists of servers, network equipment and system software with 10-20 years life-cycle and timely support. The reasons to create IT-infrastructure with the same principles are costs minimization and simplification of support. The following points that underlie during designing are high availability, flexibility and low cost. First is achieved through redundancy of hardware - doubling of servers, disks and network interconnections. Flexibility is caused by extensive use of virtualization that allows easy migration from one hardware to another in case of fault and gives users an ability to use custom system environment. Low cost – from equipment unification and minimizing proprietary solutions.

INTRODUCTION

Two BINP colliders VEPP-4M and VEPP-2000 were commissioned with feeding from VEPP-5 injection complex in 2016 [1-4]. In order to ensure continuous operation it was proposed to create highly available IT-infrastructure for both colliders and injection complex. Injection complex and collider control systems should exchange online operation data and send queries over common network to request beam and obtain beam parameters. Since all facilities have similar requirements for IT-infrastructure we tried to reduce initial deployment efforts and costs by using the same hardware and software basis for all facilities.

High availability and simple maintenance of the IT-infrastructure are the main ideas that were laid down in the base of the system. When designing it, such important requirements also where considered:

- high fault tolerance;
- recovery after failure;
- administration ease (applying specialized high-level tools);
- flexibility according to custom needs;
- equipment unification with the purpose of reducing the nomenclature (backup equipment minimization);
- the possibility of the hardware upgrading.

DESIGN DECISIONS

Hardware

After a long period of using PCs as management servers at BINP, it became obvious that their relatively low cost cannot be a decisive argument in favour of choosing this equipment for the control systems of large experimental facilities.

Personal computers have a number of disadvantages, such as:

- relatively low quality of manufacturing;
- they are not designed for continuous operation (24x365);
- poor cooling;
- limited resources;
- complexity of operational maintenance and expansion;
- inconvenient form-factor (there is no possibility to install them in telecommunication and server cabinets and racks);
- lack of advanced hardware solutions (communication interfaces, reliability enhancement technologies, etc.).

Only relatively low cost and off-the-shelf availability can be considered as advantages.

Due to foresaid, it was decided to switch to specialized server hardware, designed for long-term operation and satisfying the technical requirements for productivity in order to improve the quality of work and accessibility of the management system. As a result, the common IT-infrastructure was developed.

In general it consists of few pairs of servers. Any server in each pair can fully (and automatically) perform all the functions assigned to their pair, in case the second machine fails. Also the structure includes workstations for interaction with operators and a set of thin clients located in the most needed places on facility for local operation. To increase maintainability and availability, some of servers have redundant power supply units. Two Uninterruptible Power Supplies (UPS) are used to provide guaranteed power for all infrastructure equipment located in the rack. Each of the servers in the pair is connected to different UPS. In case of redundant power supplies in server each power supply connects two different UPS. Finally, for better availability each power supply plugged in different power lines. The scheme of hardware and network interconnection is shown on the Fig. 1.

To increase *reliability* servers are equipped with ECC memory. That can correct small errors and detect bigger ones. For remote control all servers have Intelligent Platform Management Interface (IPMI) with dedicated Ethernet port. It allows to perform any operation with server beginning from remote graphical console (full replacement for local VGA output) and finishing with BIOS configuring and firmware upgrading.

* P.B.Cheblakov@inp.nsk.su

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-THPHA048

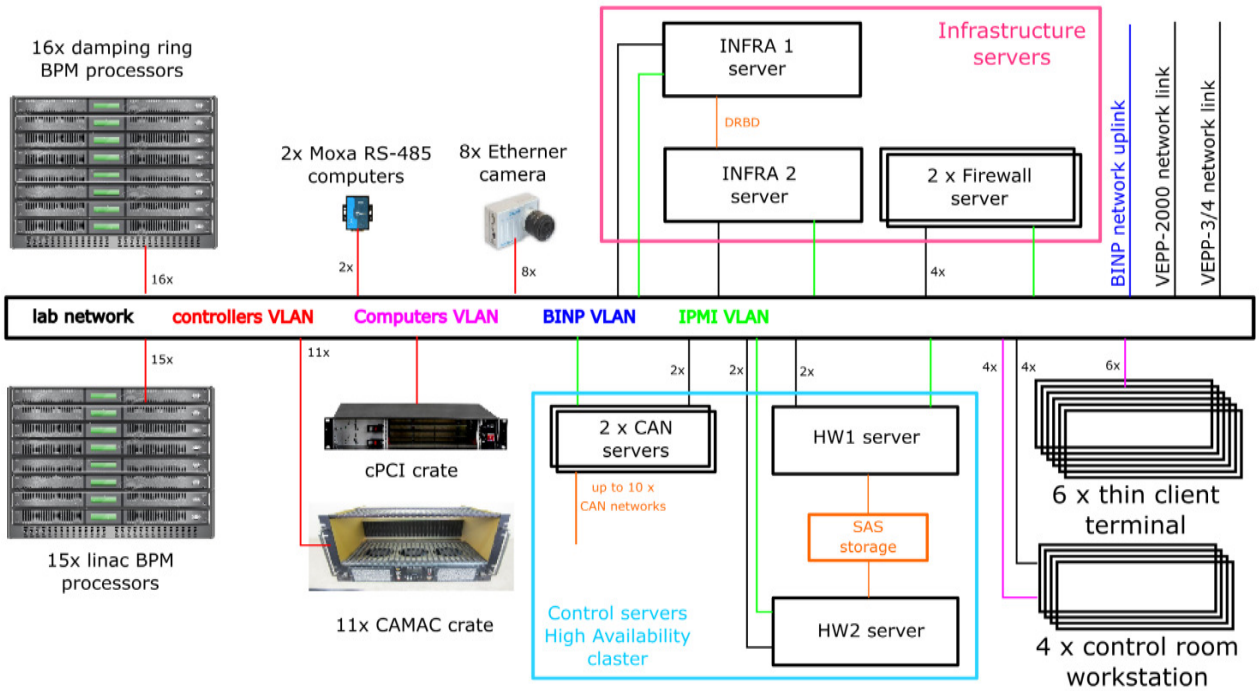

Figure 1: Injection Complex control hardware and servers.

During market analysis we have chosen server hardware from Supermicro vendor. The crucial factors was a wide offer of ready-made server platforms which are able to satisfy the most sophisticated requirements and possibility to construct custom platform from available components like motherboards and chassis.

Local Area Network

At the Injection Complex the local network central core is a stack of three HPE 1950-24 switches connected in a 10-Gb ring by means of SFP+ direct-connect cables. As a result, the core has 72 pcs 1000BASE-T ports and 8 pcs 10GBASE-T ports. The core of the network and the server are located in the same rack, which makes it easy to connect servers with 10GBASE-T ports to the switches.

In the network, in addition to the core, there is a set of 4 peripheral switches HPE 1920-24G and 10 HPE 1920-8G, which connects control-measuring equipment to the local network. On the Fig. 1 the stacked switch and peripheral switches are shown as long black rectangle across the picture.

The two main network technologies are used in building the network: VLAN and Link Aggregation (LACP). VLAN allows split a single physical network space into segments at the link layer (layer 2 of the OSI/ISO model). This approach reduces broadcast domains, which is important for low-power controllers with a simplified network stack. Also isolation of network members increases safety and stability of work. Link aggregation allows two or more communication lines to be combined into a single logical line with a larger throughput. Mentioned method is used both to implement communication between the switches, and between the switch and the server, if it has more than one Ethernet port.

At the server side Open vSwitch [5] is used for communication between virtual machines inside a host and is responsible for external communication with a switch. Open vSwitch is a production-quality, multilayer virtual switch. It is designed to support transparent distribution across multiple physical servers by enabling creation of cross-server switches irrespecting the underlying server architecture. In addition, it supports many standard management interfaces and protocols include LACP and VLAN.

Current logical structure of the entire network is shown on Fig. 2. Each accelerator has its own set of networks (VLAN) where all equipment divided by classes, e.g. broadcast controllers, unicast controllers, all management hardware (switches, UPS, IPMI etc.). Routing between these VLAN is organised by stacked switch.

At the same time each facility is isolated from others and from the Institute Network. Access to Institute Network and Internet is available only for selected hosts by means of Network Address Translation (NAT). This is done for security reason and to minimize unintentional influence.

Data Storage System

Data storage system is an important part of IT-infrastructure due to it is responsible for data availability and has a significant impact to overall performance of IT-infrastructure.

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-THPHA048

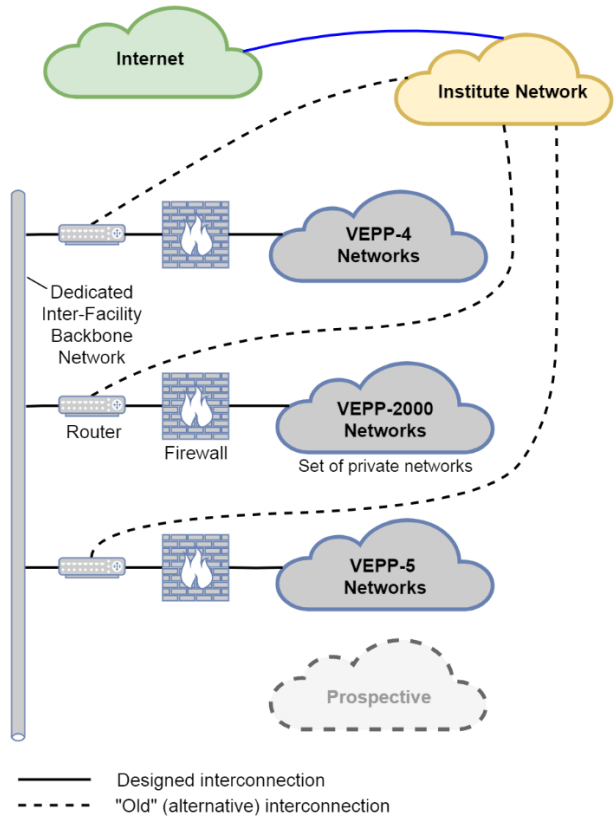

Figure 2: BINP Accelerators' networks.

Now the market presents a large variety of data storage systems, both network (Network Attached Storage (NAS)/Storage Area Network (SAN) and local Direct Attached Storage (DAS) with various access protocols. Despite the fact that ready-made systems provide many solutions for expansion, high availability, backup and automation of administration, it was decided to construct shared DAS storage system (where the same disk space is physically available simultaneously on two servers) due to a number of factors:

- SAN is very expensive and irrelevant for our tasks;
- NAS is acceptable choice but throughput is limited for all nodes by Ethernet throughput;
- DAS is the relatively inexpensive and has good performance;
- very few vendors provide shared DAS storages;
- the use of proprietary technologies can complicate maintenance and recovery after failures;
- relatively high cost of commercial systems.

It is quite important that storage system could be shared between nodes to make it possible live-migration of virtual machines. That is why we designed storage system as shared DAS.

As a result, the DAS have a 2U JBOD chassis with 12 hot-swappable HDD bays for 3.5" SAS3/SATA3 drives and an integrated IPMI module for power management and monitoring. In addition, there is a backplane with two SAS3-expanders in this chassis, which allows simultaneous connecting each of the installed disks to two servers via the SAS3-interface (see Fig. 3).

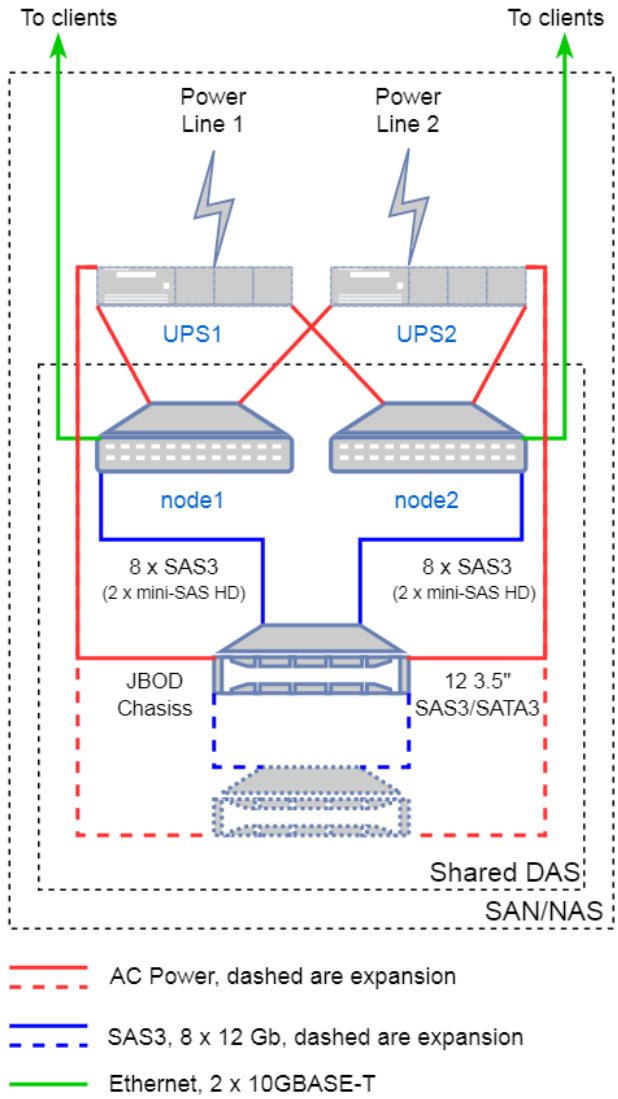

Figure 3: Scheme of the developed Data Storage System.

All the rest necessary functionality, including access arbitration of both servers to disks, is realized with the help of open software technologies and platforms, such as cLVM and Proxmox.

The JBOD chassis is connected to two main control servers (HW1 and HW2). Two firewall servers have own SSD, while other computers considered to be diskless.

Two servers (INFRA1 and INFRA2 on Fig. 1) which are responsible for infrastructure services like DNS, DHCP, network booting, user's NFS home folders, user management and etc. have shared storage based on Distributed Replicated Block Device version 9 (DRBD-9) [6]. DRBD is a software-based, shared-nothing, replicated storage solution mirroring the content of block devices (hard disks, partitions, logical volumes etc.) between hosts. DRBD mirrors data in real time, transparently and synchronously or asynchronously. As a result it functioning like shared DAS with RAID1 connected to two cluster nodes. In comparison with dedicated shared DAS it costs much less but has smaller expansion availability.

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-THPHA048

Clustering and Virtualization

Life-cycle of large experimental facility can last up to several tens of years. During this period technologies, which are used in control systems, e.g. control electronics and computers, may be changed many times. But software like operation system, libraries and applications are more changeable. And it is very likely that after some time incompatibility between some application and operation system or between operation system and hardware may arise.

To deal with that problem a virtualization technology was chosen. It allows enhance manageability, backing up data, resource consolidation and availability. We selected Proxmox Virtual Environment [7] as virtualization platform. Proxmox VE is an open source server virtualization management solution based on Kernel-based Virtual Machine (KVM) [8] and container-based virtualization with Linux Containers (LXC) [9], and includes strong high-availability (HA) support. Thanks to the unique multi-master design there is no need for an additional management server thus saving resources and allowing high availability without single point of failures. Proxmox VE includes a Web console and command-line tools, supports local and network storage types. When deployed on HA cluster with shared storage, live virtual machines can be moved from one physical host to another without downtime.

CONCLUSION

A modern hardware and software technologies like virtualization and high-availability clustering gives a new possibilities for IT-infrastructure of a large experimental facility.

Proposed and implemented IT-infrastructure scheme significantly simplify administration efforts and increase control system capabilities, availability and flexibility.

REFERENCES

[1] D. Berkaev *et al.*, "VEPP-5 Injection Complex: two colliders operation experience", in *Proc. IPAC'17*, Copenhagen, Denmark, May 2017, paper WEP1K026.
[2] F. Emanov *et al.*, "Feeding BINP Colliders with the new VEPP-5 Injection Complex", in *Proc. RuPAC'16*, Saint-Petersburg, Russia, Nov. 2016, paper WEXMH01.
[3] Yu. Rogovsky *et al.*, "Recommissioning and perspectives of VEPP-2000 complex", in *Proc. RuPAC2016*, Saint-Petersburg, Russia, Nov. 2016, paper TUYMH03.
[4] V. E. Blinov, *et al.*, "The status of VEPP-4", *Physics of Particles and Nuclei Letters*, Volume 11, Issue 5, pp 620-631, September 2014.
[5] Open vSwitch, http://openvswitch.org/.
[6] The DRBD9 User's Guide, https://docs.linbit.com/docs/users-guide-9.0/.
[7] Proxmox VE, https://pve.proxmox.com/.
[8] Kernel Virtual Machine, https://www.linux-kvm.org/.
[9] Linux Containers, https://linuxcontainers.org/.