PRESENT STATUS OF VEPP-5 INJECTION COMPLEX IT INFRASTRUCTURE

F. A. Emanov*, D.Yu. Bolkhovityanov, P.B. Cheblakov, Ya.V. Pisarev,
The Budker institute of nuclear physics, 630090 Novosibirsk, Russia

Abstract

VEPP-5 injection complex (IC) is an electron and positron beam source for VEPP-4 and VEPP-2000 experimental facilities. Continuous IC operation requires its control system infrastructure to provide high availability to all services. In order to make high availability possible and increase flexibility of control system, networks and servers were built from scratch, modern virtualization technologies being applied. The injection complex control system infrastructure is based on OS Linux and other open source software. The paper presents the architecture and implementation of the injection complex computer infrastructure.

INTRODUCTION

VEPP-5 injection complex (IC) [1–3] is an electron and positron beam source for BINP colliders VEPP-4 [4] and VEPP-2000 [5]. Layout of injection complex with colliders is shown on Fig. 1.

Since IC is a part of few continuous experiments care should be taken to avoid or reduce downtime. One of required steps is to build highly available control system infrastructure with simple maintenance for all involved accelerator facilities.

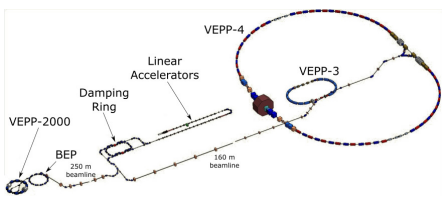

Figure 1: Injection complex and colliders layout.

Most of IC control system devices are connected to Ethernet and CAN networks [6]. CAN is mainly used at magnetic system power supply control. Modern or performance demanding devices like BPM processors, fast ADCs, embedded computers/controllers usually use Ethernet. High availability is required for server applications controlling Ethernet and CAN devices since their state is rapidly changing during routine operation.

In order to provide high availability and reduce management and maintenance efforts we developed IT infrastructure for IC and collider facilities using the same hardware and software basis [7]. Proposed infrastructure project includes redundancy for servers, server power, server network connections, storage devices and CAN connections. Injection complex control system infrastructure has been rebuilt by the moment and is able now to provide high availability for all server applications including CAN hardware control.

* f.a.emanov@inp.nsk.su

INFRASTRUCTURE DESCRIPTION

Injection complex control system currently includes 80 Ethernet and 128 CAN end devices, 8 servers, 3 control room workstations and 6 service room terminals. All servers are joined to virtualization environment cluster which is able to provide high availability for virtual machines. In order to have backup hardware for any particular task servers are organized in pairs. A number of pairs and its roles was selected to conveniently distribute services over computers

Each server of a pair has the same hardware and connections and is able to run all the pair tasks. Servers Rack has two UPSs. Some servers have redundant power supplies which are plugged in different power sources. In other cases at least paired servers have different power sources.

Network and infrastructure services are deployed to "infra" servers. Outside services, NAT and firewall are hosted on "firewall" servers. Database, control services, development machines are located on "HW" servers. In order to control CAN devices two servers with 6 PCI/PCI-E CAN adapters (each having two lines) are used. CAN lines have been rebuilt by the moment and now all devices are connected to 8 lines. We have chosen hardware from Supermicro vendor during server solution analysis. The resulting injection complex software structure is shown on Fig. 2.

In order to ensure reliable operation of control devices they should have separated network from general purpose computers. Also separation is required for management interface network and outside networks. IC network has been rebuilt by the moment using stack of three HPE 1950-24 switches as core and HPE 1920 series switches as peripheral. Network separation is implemented with VLANs. HPE 1950-24 switches support true stacking over 10GB ports which is used for server connection failover in case single switch is out of service.

HA Cluster Setup

Injection complex HA cluster is powered by Proxmox Virtualization Environment. Proxmox VE [8] is an open source server virtualization management solution based on QEMU/KVM [9] and LXC [10]. Injection Complex control system uses Centos 7 Linux for most applications thus we usually run it in KVM machines. In some special cases LXC containers are used. We use Open vSwitch as bridge for virtual machines [11]. Each server has 2 or more Ethernet ports which are connected to different central switches and bonded with LACP to make automatic failover and increase

ISBN 978-3-95450-197-7

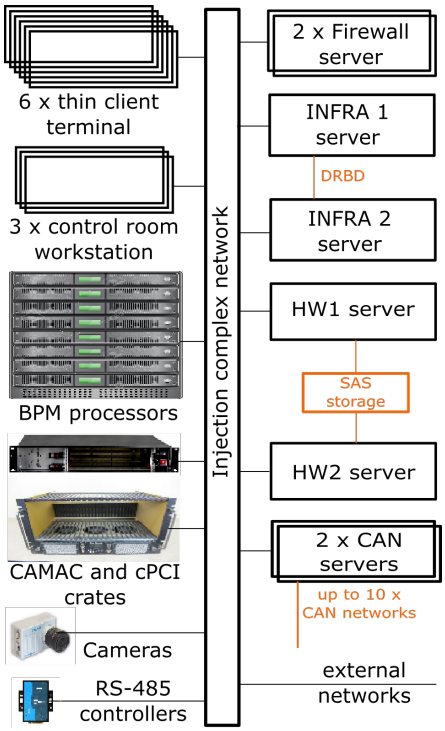

Figure 2: Injection complex IT infrastructure layout.

bandwidth. In case of shared IPMI port (CAN servers) it does not work with LACP and we use SLB balancing on open vSwitch. We tried to network-boot Proxmox VE and didn't find a way to bond Ethernet port used for boot, therefore all servers have disks with Proxmox VE installed.

Shared Storage

In order to create high-availability cluster, reliable shared storage is required. First we planned to use DRBD [12] shared storage on "Infra" servers and share it with other hosts over NFS or iSCSI for small virtual machine storage. And DAS JBOD is connected to HW servers by host bus adapters as database storage and other applications which require high capacity storage.

It is required to create redundant array from JBOD disks with LVM and share it between connected hosts for live migration. Redundancy can be created by LVM or mdadm or ZFS. Proxmox VE does not support shared LVM directly but can use LVM on iSCSI as shared. Also we can try to use CLVM which is not covered by Proxmox VE manual. We tried all these possibilities and found:

- mdadm does not work correctly with cluster.
- CLVM does not support mirroring, therefore there is no redundancy.
- ZFS is not clustered out of the box, and there is no easy way to make it clustered on Proxmox VE.
- LVM itself or nested with CLVM isn't good for live migration due to locking used in LVM.

Thus we have not found acceptable direct way to build redundant shared storage from DAS JBOD. There is indirect way: first, we created small DRBD shared storage on "hw" nodes. Then we created VM using DRBD storage, passthrough disks into this VM and created RAID10 with mdraid and shared it over iSCSI. We implemented and successfully tested this method for correct work but still need to make performance tests.

Services

In order to reduce infrastructure management efforts we deployed common set of network services like DNS, DHCP, NTP, NFS, OpenVPN, nginx webserver. In addition freeIPA is used for centralized identity management since it significantly simplifies user management. Network boot is provided by TFTP and NFS servers.

Two graphical virtual machines are used as development platform with access to accelerator environment. This machines are accessible by X2GO and also serve as server for service room terminals and other thin clients.

In order to automatize maintenance we are now deploying Ansible, which is now working on service room terminals.

Infrastructure and accelerator control services are deployed with emphasis to convenient separation between virtual machines but computer resource overhead is also taken in to account. Currently we use 21 virtual machines and containers.

CAN SERVERS REDUNDANCY

Since we try to reduce control system downtime, an automatic recovery of CAN control services is required. We believe the best way is to virtualize CAN hardware and use Proxmox cluster HA services to relocate machines in case of CAN server failure. There is no live migration for VM with any hardware passthrough but HA service can relocate such machine to host with the same hardware. In order to make it possible CAN servers have the same hardware installed and each CAN-line connected to the same ports on both servers. First we tried to passthrough CAN-adapters to KVM machine but unfortunately our hardware was not able to give all the adapters to VM. In case of LXC container it is possible to pass CAN network devices from host but it's required to install drivers for CAN adapters to hosts and modify network interface up scripts for CAN interfaces. In case of single server failure HA services will wait for watchdog. Therefore possible downtime of accelerator control services is about 1 minute.

CONTROL ROOM WORKSTATIONS

Three equal diskless workstations are used in injection complex control room. Each workstation has the same motherboard as CAN servers, with two Nvidia Quadro P2000 cards installed and 6 monitors connected. These workstations need the same set of software installed, therefore OverlayFS can be applied to reduce required storage space and simplify management. Since Linux kernel version 4.16 OverlayFS content can be exported by NFS. So virtual machine with quite a new kernel was created. This machine uses the same base layer and individual second layer of OverlayFS

ISBN 978-3-95450-197-7

and export is as root file systems for workstations as shown on Fig. 3.

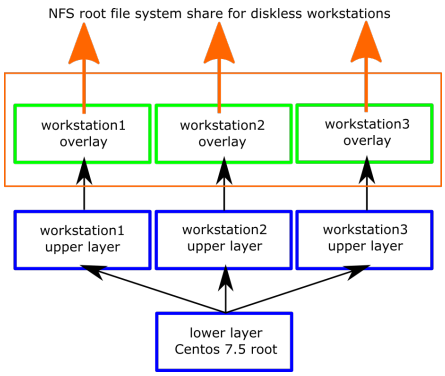

Figure 3: OverlayFS application.

CONCLUSION

VEPP-5 Injection complex control system infrastructure created by the moment involves modern network, virtualisation and software technologies in order to increase control system availability and reduces efforts for deployment, maintenance and administration.

Some infrastructure parts like DAS JBOD based shared storage and management tools are still under development.

Since we successfully deployed most of designed infrastructure for injection complex we started to distribute good solutions to VEPP-2000 and VEPP-4 control systems.

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.A. Emanov et al., Feeding binp colliders with the new VEPP-5 injection complex, Proceedings of RuPAC2016, St. Petersburg, Russia, WEXMH01.

[3] K.V. Astrelina et al., Production of intence positron beams at vepp-5 injection complex, JETP 2008, vol. 106, issue 1, pp 77-93.

[4] P. A. Piminov "Status of the Electron-Positron Collider VEPP-4", Proceedings of IPAC'17, Copenhagen, Denmark

[5] Yu. Shatunov et al., "Project of a New ElectronPositron Collider VEPP-2000," EPAC'2000, Vienna, Austria, p.439

[6] F. Emanov et al., "Present status of VEPP-5 injection complex control system", Proceedings of RuPAC2016, St. Petersburg, Russia, paper THPSC085

[7] New IT-infrastructure of accelerators at BINP, P.B. Cheblakov, F.A. Emanov, D.Yu. Bolkhovityanov, ICALEPCS2017, paper THPHA048

[8] Proxmox VE, `https://pve.proxmox.com/`

[9] Kernel Virtual Machine, `https://www.linux-kvm.org/`

[10] Linux Containers, `https://linuxcontainers.org/`

[11] Open vSwitch, `http://openvswitch.org`

[12] The DRBD9 User's Guide, `https://docs.linbit.com/docs/users-guide-9.0/`

ISBN 978-3-95450-197-7