IBM
Hybrid Cloud Integration Test Report 1Q23

IBM Hybrid Cloud Integration Test team
Jay Brenneman - rjbrenn@us.ibm.com
March 31, 2023

IBM Hybrid Cloud Integration Test (HCIT) is an effort within IBM Systems to build and maintain a customer like software environment running multiple application stacks on the LinuxONE server platform. The objective is to find IBM and open-source software defects as well as LinuxONE firmware and IBM Storage firmware defects which only can be found when running a complex software configuration over a long stretch of time. This now includes software stacks running on top of RedHat OpenShift Container Platform (OCP) as well as some potential Independent Software Vendor products from time to time.

The goal of this document is to provide a roadmap to combinations of hardware, hypervisor, operating system, and software packages that work well together. Absence from this document does not imply that something does not work with something else or is not supported by IBM, RedHat, or an ISV - it just means we personally did not try it yet.

(note: if this format looks familiar – that's because it was blatantly stolen from our z/OS peers at https://public.dhe.ibm.com/systems/z/servicetest/RSU2203.pdf)

HCIT Hardware

3931 – LA1 (A141)	Microcode at driver 51C + MCL bundle S13 released 11/16/2022 • 59 LPARs defined • 82 IFLs, 6144 GB Memory • 4 LPARs running z/VM 7.3 • 27 LPARs running Linux, various releases
3906 – M05 (M309)	Microcode at driver 36C + MCL bundle S76 released 08/03/2022 • 52 LPARs defined • 10 CPs, 100 IFLs, 18944 GB Memory • 1 LPAR running z/OS 2.4 • 4 LPARs running z/VM 7.3 • 31 LPARs running Linux, various releases

3931 – A01 (A91)	Microcode at driver 51C + MCL bundle S16+ released 03/06/2023+ • 22 LPARs defined • 98 CP, 16 IFLs, 19456 GB Memory • 1 LPAR running z/VM 7.3 • 4 LPARs running Linux, various releases • Shared with z/OS PET team
8651 – T01 (T91)	Microcode at driver 41C + MCL bundle S54+ released 01/14/2022 • 31 LPARs defined • 83 CP, 16 IFLs, 25600 GB Memory • 1 LPAR running z/VM 7.3 • 4 LPARs running Linux, various releases • Shared with z/OS PET team
3906 – M05 (M91)	Microcode at driver 36C + MCL Bundle S76 released 08/03/2022 • 26 LPARs defined • 97 CPs, 16 IFLs, 28160 GB Memory • 1 LPAR running z/VM 7.3 • 4 LPARs running Linux, various releases • Shared with z/OS PET team
DS8950F (SQ81) DS8870 (SQ54) XIV Gen3 (XIVfp1) SVC DH8 (SVCFP1)	Microcode level is R9.3 bundle 89.32.37.0 Microcode level is R7.5 bundle 87.52.45.5120 Microcode level is 11.6.2.d Microcode level is Version 8.5.0.4 (build 157.12.2206230937000)

Hardware Notes:

- A141 JEMT was replaced in January 2023, the new machine was delivered with bundle S13
- M309 and A141 are both PR/SM mode machines, with IO configuration managed by z/OS HCD from M309. There are both SR and LR coupling links from the M309 z14 to the A141 LinuxONE used to create a STP timing network for the two systems.
- M91, T91 and A91 are z/OS Platform Enablement Test machines primarily, configured by and for the z/OS team using z/OS HCD. HCIT uses a portion of those systems for performing some of our Linux work and to perform interop testing in collaboration with the z/OS PET team.

HCIT Hypervisor and Infrastructure Management Configurations

z/VM Cluster CS1	3 member SSI, z/VM 7.3 Service Level 2201 • Operations Manager 1.6 • Tape & Backup Manager 1.3 • Processors M91, T91, A91

z/VM Cluster CS2	8 member SSI, z/VM 7.3 Service Level 2201 • Operations Manager 1.6 • Tape & Backup Manager 1.3 • Processors M309, A141
RHEL 8 petKVM	4 Systems RHEL 8.6 • Kernel 4.18.0-372.19.1.el8_6.s390x • Qemu-kvm 6.2.0-11.module+el8.6.0+15668+464a1f31.2 • Processors M91, T91, A91, M309
ICIC CS1 IBM Cloud Infrastructure Center	version 1.1.6.0, 3 compute on CS1
ICIC CS2 IBM Cloud Infrastructure Center	version 1.1.5.0, 4 compute on CS2
ICIC petKVM IBM Cloud Infrastructure Center	Version 1.1.5.0, 4 compute on petKVM

Hypervisor and CIC notes:

CS1 is the 'test-ish' VM cluster, and generally will be upgraded to new releases first.

HCIT Software Configurations

Spectrum Scale fpstoc1	5 IO Servers (LPARs) 5.1.6.0 • SQ81 & XIVfp1 SAN disk • 2 CES nodes NFSv4 exports • 2 GUI & perf monitor nodes • 2 ppc64le NSD Client nodes
OCP1 cluster	Sandbox 4.11.22 2 Compute on M91, 2 Compute on T91, 2 Compute on A91 • Manually installed, has access to internet • Kubertnetes NMState operator 4.11 • IBM CSI Block driver 1.11 (auto update) • ~~IBM CNSA 5.1.5.0 mounting fpstoc1 fs~~ • IBM Storage Fusion 2.4.0 • RH OpenShift Data Foundation 4.11.6 • RH Local Storage Operator 4.11 • RH Elastic operator 5.6.3 reinstalled • RH Logging operator 5.6.3 reinstalled • RH Advanced Cluster Security 3.74.1

OCP2 cluster	Sandbox 4.10.52 6 Compute on M309 • Manually installed, has access to internet • Kubernetes NMState operator 4.10 • IBM CSI Block driver 1.11 (auto update) • ~~IBM CNSA 5.1.5.0 mounting fpstoc1 fs~~ uninstalled • IBM Storage Fusion 2.4.0 • RH OpenShift Data Foundation 4.10.11 • RH Local Storage Operator 4.10 • ~~Sysdig Security Agents 12.8.1~~ uninstalled
OCP Test0 cluster	Test 4.10.16 3 Compute on M91 z/VM • installed via ICIC ansible UPI, no internet access

Software notes:

The Fpstoc1 spectrum scale cluster runs a mixture of RHEL and SLES operating systems. The goal is to run the current and previous supported major releases. At the moment this is RHEL 8, RHEL 7, SLES 15, and SLES 12. We try to put the newer releases on the newer hardware platforms but that's more of a guideline than a rule.

Sandboxes OCP1 and OCP2 are our 'what the heck does this do?' clusters which have access to the internet. OCP1 is the 'test-ier' of the two, and may update to each new minor release of OCP as it becomes available (4.10, 4.11, 4.12..) while OCP2 will stay on the LTS releases (4.10 , presumably 4.12, etc)

Test cluster OCP Test0 is our test cluster for verifying processes to maintain OCP configurations which do not have access to the internet, as well as verifying our ansible based cluster management automation.

What service was installed:

Product	**Maintenance Level**
Z Firmware	A91 bundle S13+ -> bundle S16+
Storage Firmware	SQ81 R 9.3 89.30.68.0 to R 9.3 fp2 89.32.37.0
z/VM	CS1 no changes CS2 no changes
KVM	RHEL 8 No Changes
CIC CS1	1.1.5 -> 1.1.6
CIC petKVM	1.1.5 No Changes
OCP1	4.10.43 -> 4.11.22 in one step
OCP2	4.10.43 -> 4.10.52 in one step
OCP Test0	New install 4.10.16

NMState operator	4.10 -> 4.11 on OCP1 4.10 no change on OCP2
IBM CSI Block Driver	1.10 -> 1.11 auto updated at GA
IBM CNSA Container Native Storage Access	5.1.5.0 on OCP1 uninstalled to prep for Fusion managed 5.1.5.0 on OCP2 uninstalled to prep for Fusion managed
IBM zSCC Z Security and Compliance Center	Multiple installs 1.1.0.3 , 1.1.0.4, currently 1.1.0.6
Storage Cluster fpstoc1 GPFS	5.1.6.0 No Changes
Sysdig Security agents	~~12.8.1~~ removed to prep for RH Advanced Cluster Security
Elastic and Logging operators	New Install 5.6.3

Roll out of 1Q23 service

Our Z technicians applied bundle updates to A91. The update for the z16 was purely for currency reasons. The bundle was installed concurrently. Our A141 JEMT machine was replaced with a new one from manufacturing, and it was delivered with the current bundle at time of delivery. No problems have been noted with either machine as a result of the refreshed cde levels.

We scheduled an update of our DS8K firmware on SQ81 our DS8950f system. We used the IBM DS8000 Remote Storage Upgrade service who connected to the system over the network at the scheduled time to carefully stage new code into the machine, verified it was ready for updating, and concurrently upgraded the machine for us. The R9.3 FP2 code level was highly recommended by storage development for our pending Safeguard Copy usage. No issues have been noted with the updated code level.

We received a set of IBM Storage FS9500 units which will replace our existing SVC DH8 nodes as our ‘open system’ storage solution. This will be documented in next quarters experience report.

We added 2 members to our CS2 z/VM SSI cluster, bringing us to the maximum of 8 members there. Our current CS2 configuration places 4 members on our z14 M309 and 4 members on our LinuxONE Emperor 4 A141. We added the additional members so we have more space for additional OCP clusters.

IBM CSI Block driver auto upgraded itself to 4.11 shortly after GA. No issues were noted with the updated code level.

We upgraded OpenShift Container Platform Sandbox cluster OCP1 to 4.11 in order to prepare for the imminent release of RedHat Advanced Cluster Security support for LinuxONE. We will release an experience paper on our implementation of RH ACS in the near future. We removed the Sysdig security agents from our environment in prep for testing RH ACS.

We removed our stand alone CNSA installations from OCP1 and OCP2 to make room for Fusion managed installs of RedHat OpenShift Data Foundation as well as CNSA, and will revisit our network topology issue from 4Q22 in the future.

We installed IBM Spectrum Fusion 2.4 on both OCP1 and OCP2 environments, we then used Fusion to install RH ODF. Fusion fully automated the installation and configuration of ODF on both our sandbox systems such that it was a seamless process. We discovered that the resource request sizes of the pods deployed to run ODF are quite large, enough so that deployment is a major sizing consideration in designing an OCP cluster to support specific workloads. We discuss this later in the ‘Problems Encountered’ section.

On OCP1 we used Fusion to create a new ODF cluster and selected 3 LinuxONE internal NVMe drives which we had previously configured to 3 worker nodes in our systems. We attached the NVMe drives to the Linux virtual machines in VM with ‘DEDICATE PCIF’ statements in the z/VM Directory. No special configuration was needed to bring the devices online to CoreOS.

On OCP2 we used Fusion to create a new ODF cluster and selected 3 large DASD EAV devices which are dedicated to our worker nodes, as well as multiple virtual z/VM PAV Aliases to allow more concurrent IOPS for each EAV device. We chose to manually pre format the DASD EAV volumes with the ‘dasdfmt’ command on a separate utility Linux virtual machine so that the ODF operator would not run into timeout issues during its initialization process.

After using ‘dasdfmt’ on all three EAV devices we moved them to their permanent home and created machine config rules with ‘butane’ to place udev rules which would bring all the new devices online when CoreOS boots.

The butane file for EAV device 401 contains:

```
variant: openshift
version: 4.10.0
metadata:
  name: 99-worker-0401
  labels:
    machineconfiguration.openshift.io/role: worker
storage:
  files:
  - path: /etc/udev/rules.d/41-dasd-eckd-0.0.0401.rules
    mode: 0644
    overwrite: true
    contents:
      inline: |
        # Generated by chzdev
        ACTION=="add", SUBSYSTEM=="ccw", KERNEL=="0.0.0401", DRIVER=="dasd-eckd",
GOTO="cfg_dasd_eckd_0.0.0401"
        ACTION=="add", SUBSYSTEM=="drivers", KERNEL=="dasd-eckd", TEST=="[ccw/0.0.0401]",
GOTO="cfg_dasd_eckd_0.0.0401"
        GOTO="end_dasd_eckd_0.0.0401"
        LABEL="cfg_dasd_eckd_0.0.0401"
        ATTR{[ccw/0.0.0401]online}="1"
        LABEL="end_dasd_eckd_0.0.0401"
```

And when we use the ‘butane’ tool to compile that into something we can apply with the oc command we get the following:

```
# Generated by Butane; do not edit
apiVersion: machineconfiguration.openshift.io/v1
kind: MachineConfig
metadata:
  labels:
    machineconfiguration.openshift.io/role: worker
  name: 99-worker-0401
spec:
  config:
    ignition:
      version: 3.2.0
    storage:
      files:
        - contents:
            compression: gzip
            source:
data:;base64,H4sIAAAAAAAC/4SOQUvDQBBG7/6KZbzGuILXOSS6hGBMYHcVRCTEnVGDZQtJSWhL/3sJpW0KbXqdj/fe3IqEPTfVgkl8
L4X7WxF3N9GTTYscESoiCIR5i82HseoVEZzrIRAvSucqQwQZylA+ygcIxLNO35VGBKpaumP3P6BJYQsE9/NbDtdyuJYHZqpDTd1x045bY
69VxiLCp3P9/d73dbW3W9nT2TWLYpVNPGutXp8G535We94gHOlL8m0AAAD//wYYVhppAQAA
          mode: 420
          overwrite: true
          path: /etc/udev/rules.d/41-dasd-eckd-0.0.0401.rules
```

And we can apply the above with the usual ‘`oc apply -f <file.yaml>`’ command. We created butane and their accompanying .yaml files for the EAV devices which were at address 401 on all three worker nodes, as well as the PAV Aliases which were at addresses 4fc-4ff on all three worker nodes so that devices 401,4fc-4ff all came online automatically at CoreOS boot time. Note that the above label role specification means this configuration is applied to all workers in the cluster – even ones we did not attach a 401 device and 4fc-4ff aliases to. We are going to add more 401 devices to other worker nodes so we can scale our ODF cluster up in the future.

We reinstalled OCP cluster Test0 multiple times while developing additional tools for managing disconnected OCP clusters. We document these tools separately in our reports on managing disconnected clusters.

We reinstalled RedHat Elastic and Logging operators on OCP1 and configured them to enable external Splunk logging. This is documented separately in "Logging and Log Forwarding to Splunk in OpenShift Environment".

We installed and reinstalled multiple releases of the IBM Z Security and Compliance Center, both as a test of our new ODF based storage configuration, and to test zSCC itself collaboratively with our z/OS peers on the PET team. We are currently running zSCC 1.1.0.6 using a mix of IBM CSI Block storage and NFS dynamic storage on OCP1, and ODF CephFS storage on OCP2.

We upgraded OCP2 to 4.10.52 for currency reasons. No issues were noted yet as a result of the update.

We upgraded IBM Cloud Infrastructure Center CS1 from 1.1.5 to 1.1.6. No issues were noted yet as a result of the update.

Highlights

Deploying NFS Provisioner Software on a Disconnected OCP Cluster

We took the time to make a quick writeup that builds on our past document "Mirroring OpenShift Installation and Operator Images to a Private Registry" and describes how to add some community developed technologies into your OCP cluster. Our example is the NFS Provisioner, but the process can be followed for any available images that are built for s390x by the project team.

Clear Key Encryption with IBM WebSphere Liberty on Linux on Z

We are publishing the results of a major amount of work we have been doing to verify how to configure Linux on LinuxONE or Z systems to take advantage of the Crypto Express Accelerator adapters. Anyone wanting to run https everywhere will want to minimize the amount of cpu cycles required for all parts of the TLS negotiation process, and offloading clear key RSA algorithm processing to the CEX Accelerator card is a time tested method of doing this on LinuxONE and IBM Z systems.

When we first started looking at this last year we followed the available presentations and whitepapers, but were unable to get the desired benefit due to some changes that have happened in the Java Cryptography providers stack, particularly the IBMPKCS11Impl provider.

In a nutshell, there are now some additional configuration requirements for how to configure the pkcs11 provider in order to get the desired RSA offload ***without*** incurring undesired behavior that can cause the Java heap to run out of memory and crash or merely provide grossly unacceptable performance.

Using File-Based Catalogs with a Private Registry in a Disconnected OpenShift Environment

We have done additional work to explore how to mirror the new operator content which is file based in OCP releases 4.11 and newer vs the previous support which was SQL based. This also builds on our previous "Mirroring OpenShift Installation and Operator Images to a Private Registry" work. This document also includes a preview of installing the RedHat Advanced Cluster Security operator on LinuxONE OCP clusters, but not the entire end to end configuration of RH ACS including setting up Central.

Deploying the Petstore PostgreSQL Workload Using OpenShift Pipelines

We are beginning to build out our eventual production like test configuration automation which will include OpenShift Pipelines as part of the deployment tooling. We used the well known Petstore demo application since it pulls in the Liberty Java application server and a PostgreSQL database, as those are important runtimes and data platforms for us to ensure we are testing.

Logging and Log Forwarding to Splunk in an OpenShift Environment

We have spent some time figuring out how to get log forwarding from some of our OpenShift clusters into a Splunk demo system we have running in an available x86 OpenShift cluster. Getting it working was complex enough that we decided it needed a writeup so that you don't need to spend all the time we did getting it to work.

Problems Encountered during 1Q23

The below issues were discovered during the course of installing the new products and service during this report interval which were not fixed at the time of publication. If an issue was discovered and worked around or fully resolved it is noted how so in the Roll Out section previously.

zSCC enhancements

We test multiple beta versions of IBM Z Security and Compliance Center in collaboration with the zSCC development and test teams as well as our peers in z/OS PET integration test.

During our testing we have found some opportunities for enhancement to better support the OVN-Kubernetes networking plugin as well as adding more formal support for using multiple storage classes for the storage requiring components of the offering. We currently run all our sandbox configurations using the OVN-Kubernetes networking, and we have verified it works with multiple storage class providers.

We have tested zSCC with dynamic NFS, IBM Block Storage CSI driver, IBM CNSA Spectrum Scale, and RedHat ODF CephFS managed storage. zSCC currently only formally supports a single storage class, and the Keystore-PVC requires a RWX multi-writable volume. The IBM Block Storage CSI driver and RedHat ODF RBD cannot provide RWX volumes, but NFS, CNSA, and ODF CephFS do. It's possible to make modifications to the deployment to split the storage PVCs across two storage classes but that's kind of a hack and will only cause pain in the future as the zSCC product adds new capabilities.

VM Integrity Issue

Keep an eye out for z/VM apar VM66643 when it closes and apply it at your next planned service window to patch a system integrity issue. Benefit from our pain by applying it at your next opportunity – consult the IBM zSystems Security Portal for details once the fix is available.

ODF resource requirements

Upon deploying RedHat OpenShift Data Fabric in our environment using IBM Spectrum Fusion, we found out that ODF internal clusters do very much indeed require 30 vCPU and 72GB memory of reservation space from the workers for the first three storage devices, and 6 vCPU and 15GB of memory of reservation space from the workers for each additional set of three storage devices.

This is part of the reason why we had to remove CNSA from our configuration in the sandbox clusters. Our understanding is that this amount of capacity may be required in some circumstances during a recovery operation if the data on an OSD is lost and it is replaced with a new device. We will do some additional testing of those failure recovery scenarios in the future and write a dedicated report on the experience.

If you’re planning to deploy ODF, pay close attention to the infrastructure requirements because it absolutely requires capacity planning as a part of the overall OCP cluster configuration.

OCP worker late certificate signing request part 2 electric boogaloo

After a planned outage which left some of our OCP worker nodes down for a couple days, we brought the nodes back up and used the OCP cli to authorize new certificate requests for the nodes which had their kubelet tls certificates expire. Unfortunately, one of those requests was generated while the worker node had no functional openvswitch configuration, and it was generated with the hostname ‘localhost’ instead of ‘ocp2wrk1’. Later in that node’s startup, it began throwing error messages as it tried to find the ODF local ECKD EAV device to start the OSD service, which had moved to /dev/dasdg1 from /dev/dasdc1 in the previous boot.

We spent a lot of time trying to debug why the ODF OSD pod was using the unstable /dev/dasdXX device nodes rather than /dev/disk/by-name or /dev/disk/by-path stable disk references. It turns out that it was not an ODF or Local Discovery issue at all – the poorly created certificate was preventing the local kubelet instance from communicating with the OpenShift API service, and that was preventing the node from going into the “running” state.

As soon as we fixed the certificate issue by authorizing the backlog of certificate signing requests for that node, everything started up immediately and ODF began recovering the OSD to get it consistent with the remaining 2 storage devices in the cluster. We had not checked for additional CSRs after approving the first group immediately after restarting the down workers.

Our learned lesson is to seriously keep an eye on pending CSRs, since this is the second time it has caused us an issue. We are researching whether there is alerting in OCP we can use to monitor for this situation in the future.